Making the Grade:
The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures

Jason T. Kerwin and Rebecca L. Thornton*

January 30, 2018

Click here for the latest version of this paper

Abstract

This paper demonstrates the acute sensitivity of education program effectiveness to input choices and outcome measures, using a randomized evaluation. The program we study raises reading scores by 0.64SD and writing scores by 0.45SD. A reduced-cost version instead yields statistically-insignificant reading gains and large *negative* effects (-0.3SD) on writing. Detailed classroom observations provide evidence on the mechanisms driving the results, but mediation analyses show that observed teacher and student behaviors explain less than five percent of the differences in impacts. Machine-learning results suggest important nonlinearities and complementarities across inputs and skills in the education production function.

* Kerwin: Department of Applied Economics, University of Minnesota (jkerwin@umn.edu); Thornton: Department of Economics, University of Illinois (rebeccat@illinois.edu). We thank John DiNardo, Paul Glewwe, David Lam, Jeff Smith, Lant Pritchett, Jake Vigdor, Susan Watkins and seminar audiences at the University of Michigan, Johns Hopkins, Université Paris-Dauphine, the University of Minnesota, CSAE, Wilfrid Laurier University, CIES, the ESRC-DFID Joint Fund Poverty Conference, and London Experimental Week for their comments and suggestions. The randomized evaluation of the Northern Uganda Literacy Project would not have been possible without the collaboration of Victoria Brown, Bernadette Jerome, Benson Ocan, and other Mango Tree Educational Enterprises staff. Funding for this research was provided by the Hewlett Foundation, ESRC-DFID, an anonymous donor, and the Rackham Graduate School at the University of Michigan. All mistakes and omissions are our own. Click here to access the online appendix to the paper.

1 Introduction

Children in sub-Saharan Africa are attending school more than ever before in history – but once in school, they learn very little (Boone et al. 2013, Pritchett 2013, Piper 2010). In response, the development community has shifted away from a focus on school enrollment and towards the goal of improving learning outcomes. There is now an extensive literature examining how to achieve this goal: hundreds of studies have rigorously evaluated the effectiveness of educational interventions across a variety of contexts, countries, and types of programs.[1] A smaller but growing literature also examines the cost-effectiveness of various interventions.[2]

Yet there are important reasons to question the usefulness of this large body of evidence on "what works" in education. Systematic reviews suggest enormous heterogeneity in effectiveness across studies, making it difficult to generalize the findings from specific evaluations to inform policy (Nadel and Pritchett 2016). The variation in program effectiveness may be attributable to differences in context (e.g., India vs. Kenya) or the interventions evaluated (e.g., provision of materials vs. infrastructure upgrades), however the variation remains as large after controlling for other factors and when comparing studies that evaluate the same type of intervention (Evans and Popova 2016b, Vivalt 2017). The evidence of heterogeneity comes primarily from across-study comparisons, in part because most studies evaluate the effectiveness of a single intervention (McEwan 2015). In contrast, this paper examines the variation of intervention effectiveness within a single study – holding context and intervention type constant.

In addition to context and type of intervention, other factors affect the generalizability and policy-relevance of education program evaluations. We focus on two in this paper. First, programs studied in randomized evaluations often differ from real-world interventions in terms of design and implementation details. Because policymakers and implementers have limited resources, a common approach is to pick a highly-effective program and make it cheaper by modifying some of the most expensive inputs. This option is often appealing, since effective interventions combine

[1] Evans and Popova (2016b) discuss six systematic reviews of education program effectiveness in developing countries: Conn (2017); Glewwe et al. (2014); Kremer, Brannen, and Glennerster (2013); Krishnaratne, White, and Carpenter (2013); McEwan (2015); and Murnane and Ganimian (2014). Since their literature review, at least one additional review has been released (Glewwe and Muralidharan 2015).

[2] The evidence on cost-effectiveness is sparser due to limited data on program costs. Kremer, Brannen, and Glennerster (2013), McEwan (2015) and Dhaliwal et al. (2011) do systematic comparisons of programs' cost-effectiveness; Glewwe and Muralidharan (2015) and Krishnaratne, White, and Carpenter (2013) also discuss cost-effectiveness.

numerous inputs, many of which may seem unimportant or overly costly. However, this strategy could lead to qualitative differences in program impacts if, for example, there are important complementarities between inputs. Second, there are a large number of potential metrics that can be used to measure learning: a wide range of tests is employed, measuring a variety of skills and implemented in different languages. Moreover, the metrics chosen by researchers vary enormously across studies and may not be the same outcome that policymakers care about in a real-world setting. These variations in how learning is measured can play an important role in the interpretation of a program's measured effectiveness.

In this paper, we demonstrate how these two issues can cause misleading conclusions about how to improve learning. To do so we use a randomized evaluation of a literacy program that was conducted in 38 government schools in the Lango sub-region of northern Uganda. The program, the Northern Uganda Literacy Project (NULP), is a mother-tongue-first early-primary literacy program developed by education and curriculum experts in Uganda. The NULP involves training and supporting first- to third-grade teachers in a pedagogical strategy that focuses on teaching children to read and write in their native language, before they transition to English in grade four. The program provides material inputs including readers and primers for students, a teachers' guide with scripted steps for each lesson, classroom clocks, and writing slates. It also provides high-quality teacher training and support, including residential training and monthly classroom visits, delivered by program staff, most of whom have extensive teaching experience.

We measure the effectiveness of the NULP by randomly assigning twelve government primary schools to receive the program's entire array of high-quality inputs in their first-grade classrooms; twelve schools were assigned to the control group. The program is highly effective: after one year; it improves letter recognition by 1.01 standard deviations and improves overall reading by 0.64 standard deviations. We also find large gains in writing ability: the program improves the ability to write one's first name by 1.31 standard deviations, write one's last name by 0.92 standard deviations, and overall writing ability by 0.45 standard deviations. These reading and writing effects are comparable to some of the largest measured in the literature.[3]

Although highly effective, this program is costly for a developing-country education

[3] The confidence intervals for the reading impacts are bounded well away from zero. We can rule out effects below 0.68SD for letter names, 0.37 for overall reading, 1.03 for first-name writing, 0.78 for surname writing, and 0.17 for overall writing.

intervention, at about $15 per student per year. Scaling up the program to all students in the Lango sub-region would cost over one-quarter of that sub-region's total primary education budget. Given the infeasibly high cost of scaling up the program in its original form, a reasonable strategy would be to modify some of the elements of the program to reduce their costs. To study how reducing costly inputs would change the program's effectiveness, the NULP was modified in a way that explicitly emulated how it might be delivered at scale. Three changes were made: 1) removing the most expensive material inputs (slates and wall clocks); 2) using a cascade model of delivery where the training was conducted by government employees; and 3) providing fewer support visits to teachers. These changes amount to just a 6% difference on the Arancibia, Popova, and Evans (2016) indicators for in-service teacher training programs, yet reduce the per-student cost of the program by over 60 percent.[4] Fourteen schools were randomly assigned to receive this reduced-cost version of the program in their grade one classrooms.

While the modifications to the program were relatively small, these programmatic changes generate qualitatively different conclusions about its effectiveness. We find considerably smaller improvements in letter name knowledge in the reduced-cost version of the program than those in the full-cost program schools (0.41 standard deviations). The reduced-cost version causes no distinguishable gains when we look at more-sophisticated literacy skills (reading actual words or sentences), and gains to overall reading scores are small and statistically insignificant (0.13 standard deviations, p=0.327). The effectiveness of the two program versions diverge even further when we examine writing outcomes. The reduced-cost program shows gains only for the most basic task – the ability to write one's first name (by 0.45 standard deviations) and last name (by 0.44 standard deviations). At the same time, there are large, statistically-significant *negative* effects on the components that involved writing sentences (-0.3SD).[5]

While we are unable to conclusively identify the importance of complementarities for our results, our findings suggest that the effectiveness of an intervention can be highly sensitive to small changes in inputs. We also find that the specific outcome used to measure effectiveness matters immensely for deciding which program is more cost-effective. As measured by gains in

[4] This figure may even *overstate* the differences between the two variants, since the Arancibia, Popova, and Evans instrument is specifically designed to compare teacher training interventions (rather than all educational interventions).
[5] These findings are consistent with previous research that documents unanticipated negative consequences of education interventions (Chao et al. 2015, Fryer and Holden 2012). Both of those studies involve explicit rewards for performance. In contrast, the NULP provides no extrinsic incentives to students or teachers.

letter name knowledge, the reduced-cost version of the program is slightly more cost-effective than the full-cost version. When we look at overall reading gains, the reduced-cost version is nearly 50% less cost-effective than the original NULP.

What led to the huge success of the original version of the NULP, and why did the reduced-cost model fail? Drawing on a rich set of classroom observations and using factor analysis methods, we document important differences in classroom management and pedagogy across the three study arms. For example, teachers in the full-cost NULP are more active throughout the classroom, keeping the entire class engaged, and do fewer mass exercises on the board. We also find many commonalities: classes in both program variants are taught overwhelmingly in the local language instead of English, at significantly higher rates than in the control group.

Treating the differences in pedagogy and classroom management as independent and linear predictors of learning explains little of the difference in effectiveness between the full- and reduced-cost programs. Mediation analyses find that observable changes in behavior explain less than 4% of the difference in effectiveness across the two program variants (for both reading and writing outcomes). In contrast, a machine learning approach, which allows for higher-order terms and interactions, finds that our measured mediators can predict over 80% of the variation in classroom-average reading scores, and nearly 100% for writing scores. A set of falsification checks confirms that this is not the result of overfitting. These findings suggest that the sensitivity of the NULP's success to small changes in implementation details may be driven by powerful complementarities in the education production function.

Our findings argue for caution when modifying programs to reduce costs, adapting them to new contexts, or going to scale. Indeed, we show that taking a highly effective program and cutting down on its costs may not just make it less effective, but backfire, leaving some students *worse* off. Likewise, different learning metrics – often due to ad-hoc choices by researchers and partners – can drive vastly different conclusions about a program's effectiveness. More broadly, the degree of sensitivity of the program's effectiveness to small changes in inputs or outcome measures suggests that the current approach to evaluating education interventions may be flawed. Randomized evaluations often try to isolate the key factors that drive the success of an intervention, with the idea that we can assemble a set of effective interventions to use all at once – an inventory of "what works". This approach is likely to systematically underestimate what can be achieved by investing in education in the developing world. Our results are consistent with that story. The

massive gains in reading and writing ability caused by the NULP prove that substantial improvements in learning are possible, even in Africa's most resource-deprived schools, and using existing teachers. Education programs should take advantage of complementarities, rather than focusing on individual inputs, to address the learning crisis in Africa.

2 Context and Intervention

2.1 Context and the Northern Uganda Literacy Project

Our study is set in the Lango sub-region, an area in Uganda that is predominantly populated with speakers of a single language, Leblango; 99% of the students in our sample report speaking Leblango at home. The sub-region was devastated by civil war from 1987-2007 and suffers severe infrastructure shortages, extreme poverty and poor access to quality education. The region's schools show extremely poor learning outcomes, especially in terms of literacy. An assessment of early grade reading in 2009 found that over 80 percent of students in the Lango sub-region could not read a single word of a paragraph at the end of grade 2 (Piper, 2010).

The program we evaluate, the Northern Uganda Literacy Project (NULP), was a direct response to the poor learning outcomes in the Lango sub-region. It was developed by Mango Tree Educational Enterprises Uganda, a private, locally-owned educational tools company. Mango Tree established the NULP in collaboration with teachers, government officials, and the local Language Board. Starting in just one school, the program was piloted from 2009 to 2012 and pedagogical, curricular, and logistical refinements were made to the model to improve its effectiveness.

The NULP is multi-dimensional across both *what* is targeted to improve learning, and *how* those aspects of the education production process are targeted. Because teaching effectively in African classrooms pose multiple challenges, the model involves a carefully-designed bundle of complementary inputs. We first describe the program elements of the full program that directly address the challenges in rural Ugandan classrooms. We then describe the reduced-cost version of the program and quantify the degree to which it differs from the full-cost version using a tool developed by Arancibia, Popova, and Evans (2016).

2.2 The Full-Cost Version

Uganda's official policy is that students in early primary (grades one to three) are to be

taught in their local language before transitioning to all English instruction in grade four. In practice – however, English is still heavily used as the de facto language of instruction across the country. While it is important for students to learn English, full immersion in reading and writing a language that students do not yet know may also have powerful drawbacks.[6] The NULP was developed for one language group, Leblango, spoken by the vast majority of those living in the Lango sub-region. The program trains and supports teachers to teach literacy – reading and writing – in first grade, entirely in the students' mother tongue. Teachers are instructed not to use written English on the board or in reading materials.

Teachers in Ugandan primary schools receive their basic teacher training at primary teacher colleges across the country as well as additional training through the Teacher Development and Management System. The main training approach under this system, is a cascade model (e.g., "train-the-trainer"), in which trainers pass on skills and competences to government employees – Coordinating Centre Tutors (CCTs), who then train teachers. In contrast, the NULP provides direct training and support to teachers using experienced Mango Tree staff (expert trainers and mentors), detailed facilitators' guides, and instructional videos. Teachers undergo three intensive, residential trainings on literacy methods before each of the three terms. The first teacher training module is a five-day residential workshop on the Leblango orthography, before the beginning of the school year. In addition to the residential trainings, there are also six in-service training workshops on Saturdays throughout the year.

Under the status quo, CCTs are responsible for conducting in-service trainings and providing support to teachers through termly classroom visits. Teachers in NULP schools also receive support supervision visits conducted by Mango Tree staff members three times each term that provide teachers with detailed feedback about their teaching. In addition, CCTs are trained to provide the same type of feedback as the Mango Tree staff and given additional financial resources (for transportation and refreshments), to make two school visits per term.

Typically, teachers in Uganda rely heavily on call-and-repeat methods, where the teacher

[6] Children may simply memorize and copy words, letters, and numbers, without understanding what they are doing or how it connects to spoken words or meaning. Webley (2006) argues that education systems that use a language unfamiliar to children in school are failing. However, well-identified evidence about the causal effects of mother-tongue instruction is sparse. Rossell and Baker (2006) find that virtually all studies are focused on Spanish-language immersion in the US. The only developing-country study of mother-tongue instruction we are aware of is from Kenya: Piper, Zuilkowski, and Ong'ele (2016) found that the practice improved reading scores in one's own language by between 0.3SD and 0.6SD.

will point to a word on the board, say it, and students will repeat (Ssentanda, 2014). This call and repeat pattern can last for many minutes with a focus on memorizing whole words. In contrast, the NULP program uses a phonics-based approach, teaching students how to sound words out. Other aspects of the NULP program also help make teachers more effective. The NULP model introduces content more slowly than the standard curriculum, providing time for students to learn foundational skills. For example, only half (sixteen) of the twenty-five letters of the Leblango alphabet are taught in first grade, with the remainder taught in grade two. Teachers are also provided with scripted lesson plans for each literacy lesson.

Although schools receive a capitation grant from the government intended, in part, to pay for instructional materials (e.g., books, chalk, wall charts, teachers' guides, and resource books), the number and delivery of material resources are often inadequate. To address this, NULP classrooms are provided a set of primers (textbooks that follow the curriculum and provide visual examples for students) and readers (books that provide text for reading practice). First-grade NULP classrooms are provided with slates that allow each student to practice writing individually using pieces of chalk. The slates also allow teachers to review student writing effectively in classes of over 100 students with limited walking space (children can hold up their slates to show their work). Classrooms are also provided with a wall clock to help teachers keep track of the time during a lesson.[7]

Because the NULP provides materials, one-on-one support, and residential trainings, the model is relatively costly to implement. Not including the initial costs of curriculum and material development or broader community activities, the program costs $15.39 per student. This is more than twice the average intervention covered in McEwan (2015), and is more expensive than 93%

[7] The NULP model also involves engagement outside of the classroom. The program helps support school parent meetings once per term to discuss the importance of local- language literacy foundations, and to teach parents how to assess and support their children's literacy development at home. This involves parent training on how to interpret their child's literacy report card, and how to use a simple reading assessment tool at home. The first meeting each year also includes an activity in which children are encouraged to take a book home with them. The NULP also engages with the community more broadly through their "Strengthening a Literate Society" program to promote local-language literacy. Although the community engagement may contribute to the effectiveness of the NULP, we are unable to quantify the impact of this program element because it affects all schools in the communities we study, treatment as well as control.

of the studies in the McEwan sample.[8] Scaling the NULP up to cover all 789 primary schools in the linguistic area would cost $3.9 million per year, or over 25% of the region's entire primary education budget.[9]

2.3 The Reduced-Cost Version

Mango Tree's goal was to create the highest-quality and most-effective literacy program possible, for teachers and students in rural Ugandan classrooms. Scaling up the full-cost program, however, would be a challenge due to both budgetary and logistical constraints. Mango Tree therefore created a modified, reduced-cost version of the NULP, which was explicitly designed to resemble how the program could be implemented at scale. The inputs provided to schools in each version of the program are listed in Appendix Table A1.

There are three main differences between the full-cost and reduced-cost versions of the NULP. The first is the use of the standard cascade model of training and support, rather than working directly with teachers. This approach involved Mango Tree staff directly training government CCTs – employees of the Ministry of Education who are ordinarily tasked with training and supervising teachers in Ugandan primary schools – who would in turn train the teachers. Under the reduced-cost version of the NULP, CCTs were tasked with carrying out the teacher trainings and support visits themselves. To carry out these responsibilities, they were provided with all the NULP training materials as well as instructional videos (and solar DVD players) to show to teachers at in-service training sessions in their local communities. The second difference between the full-cost and reduced-cost versions of the NULP is that schools in the reduced-cost version received fewer support visits than those in the full-cost version: two visits per term (from the CCTs only) instead of five (two from CCTs and three from NULP staff). In both program versions, CCTs were given financial resources to make school visits and to hold training sessions. The third difference between the two versions is that classrooms in the reduced-cost version were not provided slates and wall clocks, which were seen as expensive and less-essential inputs for the program. In all, the modifications to the full program reduced the program's

[8] These figures are based on only 16 programs from Kremer, Brannen, and Glennerster (2013); the overwhelming majority of studies in the McEwan sample do not provide incremental cost information.
[9] Budget figures for 2010/11, taken from the 2014 TISSA Report (Ugandan Ministry of Education and Sports, 2014). We apportion 7% of the overall primary education budget to the Lango sub-region, corresponding with its fraction of total primary students in the country from the 2012 Uganda EMIS database.

cost by 60%, to $6.05 per student.

To further understand the differences between the two program versions, we use a set of indicators that were developed by Arancibia, Popova, and Evans (2016) to characterize in-service teacher training programs. They use their instrument to code 26 in-service training programs, including the two versions of the NULP (Appendix Table A2). Other than the total number of schools treated (12 vs. 14), there are no "overarching aspects" of the program that differ across the full-cost and reduced-cost program versions. There are also no differences in their second category, "content," between the two versions. The third category, "delivery," correctly reflects the main differences in the two program versions: direct vs. cascade training, profile of the trainers, and the number of support visits received by teachers after the initial training. In total, three indicators (5.9 percent) differ across the two versions of the NULP.[10]

The two program variants are similar in relative terms as well as absolute terms. Figure 1 presents a box plot of the differences across all pairwise comparisons of the 26 studies in the Arancibia, Popova, and Evans (2016) dataset (a total of 325 pairs). For each pair, we compute the share of indicators (out of 51) that are different, excluding three indicators related to sample size. On average, pairs of programs in the dataset differ on 53% of all indicators, compared to a difference of just 5.9% across the two NULP variants (highlighted in red in the figure). This is the lowest value across all pairwise comparisons.

3 Research Design

3.1 Sample and Randomization

The study was conducted in 76 first grade classrooms, in 38 government schools in five Coordinating Centres in the Lango sub-region of northern Uganda. Schools were eligible for the study if they met several criteria deemed important by Mango Tree to support the NULP instructional model. Using school-level data collected in late 2012, 38 schools (out of 99) met

[10] If Arancibia, Popova, and Evans (2016) had an indicator for "provision of slates", 8.33 percent of the indicators would differ across the two versions of the program.

these criteria.[11]

Schools were assigned to one of three study arms via public lottery: control schools, full-cost program schools, and reduced-cost program schools. The lottery was held at a stakeholder meeting in late December 2012, to provide Mango Tree enough time to train teachers assigned to the full-cost program schools before the start of the 2013 academic year. Prior to the lottery, schools were grouped into stratification cells – three schools in each cell – by the researchers, based on the schools' Coordinating Centre, total first-grade enrollment, and distance to the Coordinating Centre headquarters. Representatives from each school within a stratification cell drew tokens indicating treatment status from an urn.

After the second week of the 2013 academic year, enumerators collected student enrollment rosters from each school to generate an ordered list of randomly-selected students, stratified by classroom and gender. The first 25 students on the list in each of the two classrooms in a school who were present in the school on the day enumerators conducted baseline exams were selected into the sample. These 1900 first-grade students from the 38 study schools comprise our baseline sample.

Baseline tests were conducted in the third and fourth week of the school year among the baseline sample. Endline tests were conducted during the last two weeks of the school year, in November 2013; 78% of the baseline sample was tested at the endline. The longitudinal sample of 1481 students comprises our main analytic sample for the study.

Appendix Table A3 presents baseline summary statistics across each of the treatment arms, among the baseline sample, longitudinal sample, and attritors. The baseline sample is balanced in terms of basic demographics and test scores, and student characteristics also do not correlate with attrition across study arms.

3.2 Learning Outcomes

[11] The criteria were: having two first-grade classrooms and teachers, having desks and lockable cabinets for each first-grade class; having a student-to-teacher ratio of no more than 135 during the 2012 school year in grades one to three; being located less than 20 km from the coordinating centre headquarters; being accessible by road year round; and having a head teacher regarded as "engaged" by the CCT. Schools also could not have previously received Mango Tree support. Head teachers in each of the study schools were asked to assign the two best teachers in their school to their two first-grade classrooms (prior to the assignment of treatment status). In addition, prior to the assignment of schools to study arms, each head teacher signed a contract with Mango Tree outlining the guidelines for study participation. These contracts had credibility: schools that did not adhere to the contracts lost Mango Tree support in previous years while the program was piloted.

We assess student learning with exams administered at the beginning and end of the school year by trained examiners hired specifically for the testing process. Examiners were not otherwise affiliated with Mango Tree, and were blinded to the study arm assignments of the schools they visited. Exams tested first-grade students on their ability to read and write in Leblango.

Reading Leblango

We measure reading ability using the Early Grade Reading Assessment (EGRA). The EGRA is an internationally-recognized exam designed to serve as an "assessment of the first steps students take in learning to read" (RTI International, 2009). We use a version of the EGRA adapted to Leblango for use in Uganda by RTI (Piper, 2010). The exam covers six components of reading ability: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The first four components involve students attempting to identify letters, sounds, real, and invented words. The last two components have students read a simple passage aloud and then answer comprehension questions about it. The EGRA tests were conducted one-on-one by examiners sitting with individual students.

Writing Leblango

To capture students' ability to write, we use a writing assessment designed by Mango Tree to monitor writing skill acquisition. Writing tests were conducted in a group. In the first section of the test, students were asked to write their African surname and English given name. Surnames come from a small set of names that are passed down within extended families, with a known spelling in the Leblango orthography. Each name was scored separately in spelling and capitalization. In the second section of the test, students were asked to write and/or draw pictures about what they like to do with their friends. The story was scored in seven categories: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation.[12] Each writing concept is scored on a 5-point scale.[13]

[12] Presentation was added as a scoring category for endline and was not included at baseline.

[13] One of the 12 control schools was mistakenly instructed to complete the writing test in English instead of Leblango. Our results include this school, with the test marked in English. Our findings are robust to dropping the stratification cell for this school from our sample (Appendix Table A9).

Combined Exam Score Indices

The reading and writing exams consist of several modules designed to test distinct aspects of a child's ability rather than to produce a single overall score. The modules differ in their number of questions, and some are scored based on a student's speed while others are untimed. We present program effects on each module separately, as well as on combined outcome indices constructed using principal components analysis (PCA) to measure overall reading and writing ability.[14] This approach assumes that there is a single latent factor measured by each test, and that the individual components are noisy measurements of this factor. Our PCA score indices are weighted averages of the individual exam components, where the weights are the first principal component of the endline control-group data as Black and Smith (2006). We then normalize the index by subtracting the baseline control-group mean and dividing by either the baseline or endline control-group standard deviation. The endline index provides a natural interpretation: the control group mean for the index shows the control group's progress over the course of the year, and a one-unit change in the index corresponds to one standard deviation of the control-group scores on the endline exams.[15]

3.3 Classroom Observations

In addition to the baseline and endline examinations, enumerators collected classroom observations three times during the school year –July (term 2), August (term 3), and October (term 3). Each first-grade classroom in our study was observed during two 30-minute literacy lessons per visit, capturing information about classroom management, teaching style, student behavior and engagement, language of instruction, and the focus of each lesson. Classroom observations were collected in three 10-minute blocks of time using the survey instrument found in Appendix Figure A1. For each block, the enumerator indicated whether a teacher engaged in a range of pre-

[14] While there are official guidelines for scoring individual sections of the EGRA (RTI International, 2009) there is no defined system for combining the scores, although other papers that have used the EGRA to measure literacy have constructed overall scores (Aker and Ksoll, 2015). There is also no existing standard practice for producing overall writing scores.

[15] Centering the index at the baseline control-group mean, rather than the endline control-group mean, has no effect on our point estimates, since the average value of the index is absorbed by the constant in the regression. Our results are robust to an alternative index that takes the unweighted average of the normalized exam components, as in (Kling, Liebman, and Katz, 2007); we prefer the PCA index because it relates test score gains due to the treatment to the control group's progress over the year.

determined actions and recorded student actions in three categories: reading, writing, and speaking/listening. Enumerators indicated the number of minutes (out of the 10 in the block) spent on each category and the share of students participating in the activity. They then indicated whether they saw students do various actions and whether English or Leblango was used.

Our analyses involve examining the effects of the two program versions on the share of time spent on different reading and writing activities, as well as the share of time spent using the local language. We examine the specific focus of activities (i.e., sounds, sentences, words) and materials used, separately for blocks of time involving reading or writing.

In addition to analyzing the raw classroom observation variables, we also conduct factor analyses to describe classroom management and pedagogical strategies, following Glewwe, Ross, and Wydick (2016). This approach lets us summarize the patterns of correlations between different variables in the classroom observations; we then study how these patterns vary across the two program variants. We again separate our analyses by whether the block of time focused on reading or writing.

We conduct separate factor analyses for classroom management and pedagogy, retaining all factors that explain at least 10% of the variance in the data; we then apply a varimax rotation to the resulting set of selected factors (Kaiser 1958). We pool all three study arms for the factor analyses, because changes in teacher and student behavior caused by the treatment are likely to yield different factors and factor loadings. We give descriptive names to each factor based on the behaviors that load on that factor. The resulting factors and factor loadings for classroom management are shown in Appendix Table A4.[16] Appendix Table A5 shows the factors and factor loadings for reading pedagogy,[17] and Appendix Table A6 shows them for writing pedagogy.[18]

4 Empirical Strategy

[16] Three factors describe classroom management: "Keep Students Focused" comprises bringing students back on task and not ignoring off-task students, "Solid Lesson Plan" comprises referring to a teacher's guide, participating, and having a planned lesson, and "Active Throughout Classroom" comprises moving freely around the classroom, calling on individuals, and observing student performance.

[17] Five factors describe reading pedagogy: "Sounds and Letters," where students practice basic skills but not sentences; "Whole Language on Board," in which the entire class works on the chalk board on all reading elements; "Basic Elements in Breakout Sessions," which involves students working in smaller groups; "Leblango Sentences in Reader"; and "Paragraphs in Primer".

[18] Five factors describe writing pedagogy: "Pictures, Words, and Stories" (in which students use pictures as part of practicing writing words, and do not devote time to practicing letters), "Copying Teacher's Text," "Leblango Practice on Slates," "Pictures and Letters on Paper, High-Energy," and "Leblango Sentences and Handwriting".

4.1 Program Effects on Learning Outcomes

Our main outcomes of interest are student performance on reading and writing measured by the EGRA and the writing test. We estimate the effects of the NULP on each test component separately, and on overall reading and writing performance using the PCA index described above. Our empirical strategy relies on the random assignment of schools to the three study arms for identification. Randomization allows us to attribute post-treatment differences in outcomes to the effect of the program the school received because the students and teachers in the three study arms is balanced, in expectation, on observed and unobserved pre-treatment variables.[19] We run regressions of the form:

$$y_{is} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{L}_s'\gamma + \eta y_{is}^{baseline} + \epsilon_{is} \qquad (1)$$

Here i indexes students and s indexes schools. y_{is} is a student's outcome at endline — typically his or her score on a particular exam or exam component. $FullCost_s$ and $ReducedCost_s$ are indicators for the school being assigned to the full-cost and reduced-cost version of the program; the omitted category is an indicator for being assigned to the control group. ϵ_{is} is a mean-zero error term. β_1 and β_2 are our estimates of the effects of the full-cost and reduced-cost programs, respectively. We control for a vector of indicator variables for lottery stratification cells $\boldsymbol{L}_s$ in order to consistently estimate treatment effects and to improve the precision of our estimates. (Bruhn and McKenzie, 2009).[20] Our preferred specifications also control for the baseline value of the outcome variable, $y_{is}^{baseline}$, as specified in our pre-analysis plan, to address any potential baseline imbalance on test scores.[21] We also show our results without baseline controls. To account for the fact that the treatment was randomized at the school level and our outcomes are correlated

[19] Consistent estimation of β_1 and β_2 requires that the treatment indicators are independent of the error term ϵ_is once we condition on the other controls in the regression. Random assignment of schools to study arms guarantees this independence under the stable unit treatment value assumption (i.e. no spillovers). This is plausible in our context because the schools are generally too far apart for spillovers to plausibly occur. The average school in our sample is 5 km (3 miles) from the closest in a different study arm, which would mean students and teachers would have to walk an additional 50 minutes to reach the other school. The closest two schools in different study arms are 0.9 km (0.5 miles) apart, or 10 minutes walking time. These distances assume straight-line travel between the schools; actual routes are typically even longer.

[20] The probability of assignment to each study arm varies by stratification cell because two cells contained just two schools instead of three.

[21] https://www.socialscienceregistry.org/docs/analysisplan/36/document

within schools, we report robust standard errors clustered by school.

4.2 Hypothesis Testing

We conduct all hypothesis tests in this paper using randomization inference, following Athey and Imbens (2017). This approach approximates the exact *p*-value for our observed treatment effects under the sharp null hypothesis that the treatment effect is zero. It also addresses the issue that cluster-robust standard errors can be too small if the number of clusters is low (Cameron, Gelbach, and Miller 2010). The typical cutoff is 50 clusters; our study has just 38.

The randomization inference procedure consists of running simulated versions of the random lottery that was used to assign schools to study arms. Within each stratification cell, we randomly re-assign schools to study arms, and then estimate the treatment effects for these simulated assignments using equation (1). Repeating this 1000 times gives us a distribution of the treatment effects we would expect under the null hypothesis of a zero effect, where any evident treatment effects are simply due to chance. Building on software developed by Heß (2017), we modify this basic approach to account for the multiple treatment groups in our study. For each regression, we conduct three hypothesis tests – a comparison of the full-cost treatment with the control group, a comparison of the reduced-cost treatment with the control group, and a comparison of the two treatment groups with each other – by permuting only the two study arms in question. All the reported *p*-values and indications of statistical significance in this paper are based on this randomization inference procedure.

4.3 Identifying Mechanisms through Classroom Observation Data

We use the classroom observations data to explore how teacher and student behavior were affected by each version of the program via several methods: reduced-form estimation, mediation analysis, and machine learning.

Reduced-Form Estimates

To measure the reduced-form effects of the two program variants on teacher and student behavior, we use the classroom observations data at the level of a 10-minute observation block. Our regression model is:

$$y_{blrcs} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{L}_s'\gamma + \boldsymbol{R}_r'\delta + \boldsymbol{E}_{rcs}'\rho + \boldsymbol{D}_{lrcs}'\mu + \boldsymbol{B}_{blrcs}'\omega + \epsilon_{blrcs} \quad (2)$$

where *s* indexes schools, *c* indexes classrooms, *r* indexes the round of the visit, *l* indexes the lesson being observed, and *b* indexes the observation block. In addition to the variables that appear in equation 1, equation 2 adds as controls vectors of indicators for each observation round ($\boldsymbol{R}_r \in \{1,2,3\}$), enumerator ($\boldsymbol{E}_{rcs}$), day of week of the observation ($\boldsymbol{D}_{lrcs}$), and the order of the observation block ($\boldsymbol{B}_{blrcs} \in \{1,2,3\}$) within the lesson. ϵ_{blrcs} is a mean-zero error term. We cluster the standard errors by school. Our outcomes, *y*, are both the raw variables and the factor analysis indices from the classroom observations data described above.

Mediation Analysis

In addition to measuring reduced-form effects, we would like to understand how much of the program effects on test scores can be explained (mediated) by the changes in observed behavior in the classroom. Simply re-estimating the main regression equation with the classroom observation variables as controls can lead to arbitrary bias in the estimated treatment effect and does not reveal the extent to which the treatment effect is explained by the mediator (Acharya, Blackwell, and Sen 2016). Instead, we use the sequential *g*-estimator of Acharya, Blackwell, and Sen (2016) to estimate what proportion of the treatment effect is explained by mediators – variables affected by the treatment that in turn influence the main outcome.

Sequential g-estimation involves three steps. The first step is to estimate the effects of the mediators on the outcome variable. Second, use those estimates to remove the effects of the mediators from the outcome variable, creating a "demediated" outcome. Third, regress the demediated outcome on the treatment indicator to obtain the estimated effect of the treatment on the outcome, net of the changes in the mediators.

The Acharya, Blackwell, and Sen estimator is only applicable to a single binary treatment variable, so we restrict our attention to a pairwise comparison between the full-cost and reduced-cost versions of the NULP. This allows us to explore the mechanisms behind any differences in outcomes between the two program variants. We present results using the classroom management and pedagogy factors constructed from the classroom observation data as mediators, but the results

are nearly identical if we use the raw classroom observation variables instead. We re-center the mediator variables (i.e., the factor indices) relative to the reduced-cost program, by subtracting off the reduced-cost program mean. We then run the following regression to estimate the effect of the mediators on the outcome:

$$y_{isc} = \beta_0 + \beta_1 FullCost_s + \boldsymbol{M}'_{sco}\tau + FullCost_s * \boldsymbol{M}'_{sco}\lambda + \boldsymbol{I}'_{sco}\pi + \boldsymbol{L}'_s\gamma + \eta y_{isc}^{baseline} + \epsilon_{is} \quad (3)$$

The notation follows equation 1, but also includes, a vector of mediator variables, where *o* indexes a specific observation block (10-minute time period) in classroom *c* and school *s*. We allow the effect of the mediators to vary across study arms by including interaction terms, following Acharya, Blackwell, and Sen (2016).

To consistently estimate τ and λ, we need to satisfy a "no intermediate variable bias" assumption – that there are no variables omitted from our regression that are affected by the treatment and influence the outcome and also correlated with the mediators. While we cannot guarantee that we have accounted for all potential intermediate confounders, we mitigate this possibility in two ways. First, our vector $\boldsymbol{M}'_{iso}$ includes all the factor variables summarizing the classroom observations data.[22] Second, we control for a vector of intermediate variables $\boldsymbol{I}'_{iso}$ that could be confounders: fixed effects for the block of the classroom observation, the round of the visit, the day of the week, the enumerator who conducted the visit, and a control for the total number of observation blocks for a given classroom observation.[23] While our classroom observation data is extremely rich, making the "no intermediate variable bias" assumption plausible, we cannot rule out all potential violations. As a result, the findings from our mediation analysis should be taken as suggestive rather than definitive.

After estimating (3), we then construct a de-mediated value of *y* by subtracting two terms from the raw outcome (y_{isc}): 1) the product of the mediators and the estimated coefficient ($\boldsymbol{M}'_{sco}\hat{\tau}$), and 2) the product of the treatment indicator, the mediators, and the estimated interaction coefficient ($FullCost_s * \boldsymbol{M}'_{sco}\hat{\lambda}$). This yields the following expression:

[22] Because we use the factor analysis indices as our mediator variables instead of the raw classroom observation variables, it is conceivable that some of the raw variables could be intermediate confounders. However, if we instead use the full set of raw variables, our results barely change.

[23] The classroom observations were typically at least thirty minutes long and hence contain three observation blocks. In a small number of cases (5% of our mediation analysis sample) the class ended early and the number of blocks was less than three; this rate did not vary by study arm.

$$y_{isc}^{demediated} = y_{isc} - \boldsymbol{M}'_{sco}\hat{\tau} - FullCost_s * \boldsymbol{M}'_{sco}\hat{\lambda} \quad (4)$$

The result, $y_{isc}^{demediated}$, can be interpreted as the outcome variable purged of the effects of changes in the mediator variables. We then can estimate a modified equation 1, regressing the de-mediated value of *y* on the treatment indicator and our baseline controls (recall we are not using control-group observations in these analyses):

$$y_{isc}^{demediated} = \beta_0 + \beta_1 FullCost_s + \boldsymbol{L}'_s\gamma + \eta y_{is}^{baseline} + \epsilon_{is} \quad (5)$$

Acharya, Blackwell, and Sen (2016) show that under the assumption of no intermediate variable bias, equation 5 estimates the *average controlled direct effect*. In our case, this is the difference in test scores between the full-cost treatment and the reduced-cost treatment, under the counterfactual hypothesis that all mediators are held at the mean value in the reduced-cost study arm. This allows us to measure what proportion of the treatment effect can be explained through changes in the mediators we measure through the classroom observations. Specifically, we compare this estimate to the main treatment effect estimates from equation (1) to assess the share of the change in test scores driven by changes in our measured mediators.

Machine Learning

Standardized classroom observation measurements can be strong predictors of student achievement in developed countries (Kane and Staiger 2012). Recent research by Araujo et al. (2016) shows that scores from the CLASS observation tool can predict student success in developing country schools as well. Our data collection instrument differs from the CLASS tool in that it focuses on objective behaviors rather than on subjective assessments of teaching quality.

Since our tool has not been validated in other contexts, we use a machine-learning approach to assess its predictive power for endline test scores. For our machine-learning analyses, we no longer face the limitations that kept us from including all three study arms in our mediation analyses, so we work with the entire dataset. First, we collapse the data to classroom-level means and estimate the following regression:

$$\bar{y}_{sc} = f(\bar{\boldsymbol{M}}'_{sc}) + \epsilon_{sc} \quad (6)$$

Here $\bar{y}_{sc}$ is the average endline exam score in classroom *c* in school *s*, $\bar{\boldsymbol{M}}'_{sc}$ is a vector of the average values of the mediators in classroom *c* in school *s*, and $f(\,)$ is a flexible function of the classroom-average mediators. We would quickly encounter dimensionality problems if we simply estimated an OLS regression and included higher-order terms and interactions in our regression. There are only 72 classrooms in our dataset (after dropping three classrooms that cannot be linked to the classroom observation data), and we have a total of 26 mediators. Instead, we use a machine-learning approach to assess the predictive power of our mediator variables. We apply the kernel regularized least squares (KRLS) approach of Hainmueller and Hazlett (2014) to the data, allowing the estimator to select the interactions and higher-order terms. This estimator converges to the true R-squared asymptotically; we also assess the extent of overfitting.

5 Results

5.1 Learning under the Status Quo

Data from the study's control group confirms that students in the Lango sub-region have extremely poor learning outcomes. Over 80% of students entering the first grade are unable to recognize a single letter of the alphabet, and the majority of those students leave the first grade having made no progress whatsoever (Figure 2, Table 1). At the end of first grade, roughly 50% of students could recognize only a single letter of the alphabet. Just over 20% could recognize between one and five letter names, and a similar fraction could recognize between six and twenty. Fewer than 10% of pupils could correctly identify more than twenty letters out of 100 total questions in the letter grid. These results are consistent with other studies in sub-Saharan Africa, which find extremely low learning levels. (Boone et al. 2013, Pritchett 2013, Piper 2010)

Overall reading performance mirrors the performance on letter-name recognition: 40% of students get at least one correct answer across the six components of the exam at the beginning of the school year, but that number rises to just 60% by the end of the year. A small number of high-performers do much better than the typical student: the fraction of students answering more than twenty questions right rises from roughly 0% at the beginning of the year to 10% by the end of the

year (Figure 2).[24]

The measured increases in exam scores in the control group form a natural benchmark comparison for the effects of the two versions of the NULP. In the absence of the intervention, students improve by 0.15SD in reading over the course of first grade, and 0.47SD in writing. Our randomized evaluation measures the additional gains caused by the program.

5.2 Program Effects on Reading

Figure 2 Panel B shows the endline distribution of EGRA scores by study arm. The full-cost version of the NULP reduces the proportion of students who cannot answer a single question on the exam by nearly half, and more than triples the share that can answer 21 or more questions (out of 277 total questions on the exam). The reduced-cost version of the NULP achieves improvements in EGRA performance, but does so by a smaller degree than the full-cost variant.

The impacts of the two versions of the NULP on EGRA scores, estimated using equation 1, are shown in Table 2.[25] The full-cost version of the program has a very large impact on letter-name knowledge: scores increase by 1.01 standard deviations (Column 2). The reduced-cost program improves letter-name knowledge by 0.41SD, which is still a meaningful gain but less than half as the full-cost version of the program, and is not statistically significant (p=0.106). The difference between the effects in the full-cost and reduced-cost programs is 0.61 standard deviations and is statistically significant at the 0.01 level.[26]

Examining the effects of the two versions of the program on the other EGRA components reveals a more nuanced picture. The full-cost program has strong effects on all six components of reading, and five of the six effects are significant at the 0.05 level. The reduced-cost program,

[24] A notable exception to is reading comprehension, which has the highest proportion of students getting a question right at 30% – more than the share who were able to read any of the relevant passage aloud. This pattern is identical across study arms. The high scores may be because students are better able to make words out on the page than to correctly pronounce them out loud, and also may the result of lenient scoring by the examiners. They could also be due to guessing: the first question translates to "When does [the character] like to go visit grandmother?" and the correct answer is "during holidays/vacation", which may be easy to guess.

[25] The estimated effects on EGRA performance are virtually unchanged when we omit the baseline exam score controls (Appendix Table A7).

[26] The statistical significance of the main treatment effect on reading is robust to correcting for multiple comparisons. Applying the conservative Bonferonni correction would mean multiplying the p-value by 6, because we run three hypothesis tests each for both the reading and the writing index. The adjusted p-value would be 0.03. The p-value for the difference between the two NULP variants would be 0.06, and thus remain significant at the 0.10 level even under this conservative standard.

however, has no statistically-significant effect on any EGRA component.

This finding is demonstrated further with the estimates for the combined reading score index (Column 1). The full-cost program raises this index by 0.64SD, confirming that the large effect of the program is not merely an artifact of focusing on knowledge of letter names. Taking 0.64SD as the best estimate of the program's impact on reading ability, the effect of this program is among the largest ever measured in a randomized trial of an education program. The reduced-cost program's effect on the EGRA index is just 0.13SD and is statistically indistinguishable from zero.

Figure 3 illustrates the NULP's effects on reading relative to the interventions in the McEwan (2015) meta-analysis that report reading outcomes (excluding studies that combine reading with other learning outcomes).[27] The full-cost NULP program has larger impacts on reading than any study in McEwan's dataset while the reduced-cost version is closer to the average. Our estimates of effects are also much more precise than the typical high-impact program evaluation. In the few cases where randomized evaluations of education programs have found large effects, those estimates have typically been paired with wide confidence intervals that do not exclude much smaller impacts.[28] We can reject test score gains smaller than 0.37SD at the 0.05 level.

5.3 Program Effects on Writing

Columns 2 and 3 of Table 3 show that both versions of the program have large effects on students' ability to write their first and last names.[29] The full-cost program also has positive effects on students' ability to write or draw a short story (Columns 4 to 10). Altogether, the combined writing score rises by 0.45SD (Column 1), which is statistically significant at the 0.1 level.

In contrast, the reduced-cost program has uniformly negative effects on story-writing, with the negative effects on Voice and Presentation reaching significance at the 0.05 level. The combined writing score falls by 0.16SD, although this drop is not statistically significant.

[27] See the online appendix for a list of all the papers used in this figure.

[28] Two interventions in McEwan (2015) have effect sizes comparable to the impacts we estimate: Carrillo et al. (2010) cannot reject zero for one of their study arms, and the other has a 95% confidence interval ranging from 0.4 to 1.4; Tan et al. (1999) have a 95% confidence interval that ranges from about 0.12 to 0.68. Neither study reports outcomes for reading alone.

[29] Just as for reading, the writing test results are essentially unchanged if we omit the baseline exam score controls (Appendix Table A8).

However, the gap between the effects of the two program variants is statistically significant for every measure of writing ability ($p<0.05$) and quantitatively large.

Figure 4 parallels Figure 3, but for writing. It shows the effect of the two NULP variants on writing scores, compared with the interventions in the McEwan (2015) meta-analysis dataset reporting writing outcomes (excluding studies that only report writing and other learning outcomes in a combined score).[30] The full-cost program has a larger impact on writing than any of the studies in the McEwan data while the reduced-cost program has a large negative effect.

5.4 Cost-effectiveness

The large effects of the program naturally raise the question of its cost. The full-cost version of the NULP cost $15.39 per student, which is at the 93rd percentile of the cost distribution for the 16 treatments covered by McEwan (2015).[31] The reduced-cost version was designed explicitly to emulate how the program could be taken to scale. This version cost $6.05 per student, at the 63rd percentile of studies in McEwan (2015). To compare the cost-effectiveness of the two versions of the program, we present the cost per student of each program version, as well as the cost per 0.2SD gain and the SD gain per dollar spent for three different measures of the program's effectiveness (Table 4).

We first present the cost-effectiveness results using the estimated program effects on the most basic reading skill: letter-name knowledge. Using this measure, the two versions are relatively comparable, both increasing letter name knowledge by 0.07SD for each dollar spent. The full-cost program is a bit more costly per student per learning gain, costing an extra six cents per student to raise letter name knowledge by 0.2SD. In other words, the reduced-cost program could be scaled to reach nearly three times as many students for the same total expenditure as the full-cost version.

Assessing cost-effectiveness based on overall reading ability reverses this conclusion. The full-cost version of the program yields over twice the gains in performance per dollar compared to the reduced-cost version – 0.04SD per dollar vs. 0.02SD per dollar. Similarly, the cost per 0.20 standard deviations increase in reading is $4.85 in the full-cost program and $9.10 in the reduced-

[30] See the online appendix for a list of all the papers used in this figure.
[31] These figures are based on actual expenditures in 2013 and include only incremental program costs, excluding costs related to materials development, curriculum design, or the community engagement aspect of the program.

cost version.

Using effectiveness estimates from the writing ability index shows an even starker pattern: because the reduced-cost version of the program reduces writing performance, the cost per 0.2-SD gain from that version of the program is undefined. Instead, each dollar spent on the reduced-cost version of the program *decreases* writing performance by 0.03SD.

Figure 5 shows the cost-effectiveness estimates for our study as well as all studies from McEwan (2015) that contain incremental cost data.[32] Compared to other interventions, the full-cost NULP is near the high end of the distribution of cost effectiveness, irrespective of the metric used to judge cost-effectiveness: the relative cost-effectiveness of the full-cost program is at the 78th percentile of the overall distribution for writing and at the 84th percentile for both reading metrics. In contrast, the relative cost-effectiveness of the reduced-cost version of the program is highly sensitive to the outcome we choose.

6 Mechanisms

The full-cost version of the NULP has significant benefits for pupil literacy across all metrics of reading and writing. In contrast, the reduced-cost version seems to achieve gains on only the most basic outcomes – letter recognition and name writing – with no gains in other areas and statistically-significant losses on more advanced aspects of writing. The two variants of the program were randomly allocated as complete packages, so we cannot causally separate the effects of each individual input. Instead, we exploit detailed classroom observation data to understand mechanisms, examine how the two versions of the program affected teacher and student behavior in the classroom, and test the role of those changes in mediating the program effects. We then use these results and discuss potential reasons why the effectiveness of the NULP are so sensitive to minor changes in inputs and implementation. In particular, the evidence suggests that certain program components function as strong complements to one another, so removing one can lead to sharply different results.

[32] We omit two of the three outcomes from Oster and Thornton (2009). Those outcomes have (insignificant) negative effects and extremely low incremental costs, leading to large negative cost-effectiveness estimates that make the figure difficult to read. See the online appendix for a list of all the papers used in this figure.

6.1 Allocation of Time across Classroom Activities

We first look at how teachers allocate class time to different activities in the classroom. Table 5 shows the proportion of the 10-minute observation block allocated to reading, writing, and speaking/listening, and the proportion of time where the local language is used. Teachers in both versions of the program spend substantially more time on reading and much less on speaking and listening. The drop in speaking and listening time is 2.2 percentage points larger in the reduced-cost version of the program, though this difference is not statistically significant (p=0.193). Teachers in the full-cost version of the program actually spend slightly less time (3.6 percentage points less, p=0.148) on writing than the control group (Column 2). Given relative the impacts on writing, this suggests that time spent learning to write was not due to increased time on task, but rather that time spent learning to write was more productive in the full-cost program.[33]

Teachers in both versions of the program use Leblango more often in the classroom than those in the control group. The difference between the use of the local language in the full-cost and reduced-cost versions of the program is just 3.2 percentage points (and not statistically significant), although the control group already uses Leblango 69% of the time. Given the high base rate of mother tongue instruction, and the fact that the program effects are very different between the two programs, it seems unlikely that the mother-tongue focus of the NULP is a key determinant of its effectiveness.[34] However, it may be an important complement to other inputs, which is a possibility we will return to below.

6.2 Classroom Management and Pedagogy

We examine how time is allocated during reading activities in Table 6. Students in the full-cost program are more likely to spend time reading from readers and primers – materials that the

[33] Using the data on time on task and the estimated differences across study arms, we calculate that students in the full-cost program gained 0.024SD in writing scores for every hour spent learning writing, as opposed to 0.011 for the control group and 0.008 for the reduced-cost group. Time spent on reading is also much more productive in the full-cost program than in the other two study arms. Students in the full-cost program gain 0.011SD on the EGRA for each hour spent on reading, as compared with 0.004SD per hour in the reduced-cost program and 0.002SD per hour in the control group.

[34] Given the substitution from English to the use of Leblango as a result of the NULP, one question is how the program's effects spill over onto English proficiency. We find no evidence of a decline in English speaking ability in either treatment arm, and for more open-ended questions on the English speaking test there are gains of about 0.3SD for the full-cost study arm, with one of those gains being statistically significant at the 0.10 level (Appendix Table A10).

NULP provides to classrooms (Columns 5 and 6). The effect on primers is statistically significant at the 0.05 level. Although full-cost and reduced-cost classrooms both received the same primers and readers, we see a smaller effect on material use for the reduced-cost program (though the differences are not statistically significant). The control group hardly uses these materials at all – just 3% of the time for primers and 6% for readers – reflecting the extremely low availability of those types of classroom materials.

Reading activities are more likely to focus on sounds in both versions of the program, reflecting the phonics-based emphasis of the NULP. The difference between the full-cost and reduced-cost versions is statistically insignificant but non-trivial, with over 50% more focus on sounds in the full program. There are no detectible differences in practicing letters, words, and sentences across the three study arms. However, based on exam score measures, students in the full-cost program perform much better on these aspects of reading. This suggests again that the gains from the program are not simply due to additional time on task (i.e., on letters, words, and sentences), but that the time spent is more productive. Another possibility is that the control group moves too quickly through the curriculum, or does not spend enough time building a foundation of basic skills (i.e., sounds) upon which more-complicated skills can be built (Pritchett and Beatty 2015).

The impacts on the factor analysis indices for time spent reading in Table 7 reveal more subtle patterns. Classroom management during reading classes differs slightly across the study arms. Full-cost program teachers are more likely to be active throughout the classroom and reduced-cost teachers are somewhat less likely; the difference between the two versions is significant at the 0.10 level. We also see different pedagogical strategies across study arms. There is an increase in practicing reading Leblango sentences from readers; this change is larger for the full-cost classrooms but the difference is not statistically significant (Column 7). Both full- and reduced-cost program students spend more time reading paragraphs out of primers (Column 8). These activities replace whole-language exercises at the chalkboard (where the teacher covers all the different literacy concepts at once) and working on basic elements in small groups (Columns 5 and 6). The decline in whole-language exercises on the board is statistically significant for the full-cost classrooms, and the difference between the two program versions is significant at the 0.10 level.

Tables 8 and 9 show the program effects on time allocation and the use of materials during

writing activities. Students in both the full-cost and reduced-cost classes spend more time on name-writing (Table 8, Column 5); the full-cost treatment effect is 40% larger than the reduced-cost effect, but the difference is not statistically significant (p=0.199). Critically, the reduced-cost group spends significantly *less* time than the control group on writing sentences (Column 4). This is consistent with their declines in performance on the story-writing component of the writing test (Table 3).

There are large differences in the use of materials across the two program versions. Full-cost program students are much more likely to practice writing on slates, which substitute for writing on paper (Table 8, Columns 7 and 8). In contrast, reduced-cost program students spend significantly more time than full-cost program students on "air-writing" – tracing out the shapes of letters in the air (Table 8, Column 6). Full-cost program students also spend much less time copying their teacher's text, and much more writing their own text. The latter gain is absent for the reduced-cost program students, and the difference is statistically significant (p=0.002).

The classroom management and pedagogy factor analysis indices for writing in Table 9 tell a similar story. The index for practicing Leblango using slates is massively higher among full-cost program students (Column 6). It increases somewhat for reduced-cost students, reflecting the fact that this index loads on multiple underlying variables and can be positive even in the absence of slates. The estimates possibly reflect the reduced-cost teachers attempt to carry out the NULP pedagogical model, but achieve limited success because they lack a key input (slates). Both program versions show drops in copying the teacher's text; this effect is again significantly larger among full-cost students (Column 5). Full-cost students also have lower values for the index of drawing pictures and letters on paper with high levels of energy and participation (Column 7); class activities that combine pictures, words, and stories become more common instead (Column 4). Classroom management during writing classes differs significantly across study arms. Teachers in the full-cost treatment arm exhibit an increase in being active throughout the classroom (Column 3). The difference across treatment groups is itself statistically significant. This could be due to the slates, which make engagement with all of the students in full program classrooms easier.

6.3 Mediation Analysis

The reduced form results on classroom observations highlight some potential mechanisms through which the full-cost and reduced-cost NULP versions may have differentially affected

student performance in reading and writing. It is unclear, however, how much of the differences in treatment effects can be attributed to the differences in time allocation, classroom management, and pedagogy. To understand this further, we conduct a mediation analysis using a version of the sequential *g*-estimator of Acharya, Blackwell, and Sen (2016) described in Section 3.4. This analysis addresses the question: what would be the effect of the full-cost version (relative to the reduced-cost version) if the variables in the classroom observations did not change? This allows us to summarize how much of the difference in treatment effects can be explained by changes in mediators.

The results, presented in Table 10, rely on the assumption that there are no other unobserved mediators that are correlated with the ones we observe, and so should be interpreted as suggestive. With that caveat, we find that the changes in classroom observation mediators explain only a small fraction of the difference in the treatment effects across study arms: 2.0% for reading (1.1% for letter name recognition alone) and 3.7% for writing with an adjusted R-squared of 0.59 and 0.331, respectively.

The limited ability of the mediators to explain the treatment effects has several possible explanations. First, the relevant changes in teaching may be too subtle to detect using our classroom observation instrument. For example, while we can measure time use and specific actions taken by teachers and students, we are unable to determine how effective the time used and the actions taken are in terms of causing learning. As discussed above, differences in the productivity of time on task may be a crucial part of the story behind our results. Second, our classroom observation tool may simply provide a poor measurement of classroom behaviors. Third, the relevant changes in teaching may be combinations of a variety of different complementary inputs.

In the next section, we use a machine-learning approach to explore whether a functional form allowing for complementary inputs and non-linearities might better predict test-score gains than using the classroom observation variables as linear predictors.

6.4 Machine Learning

We estimate a kernel regularized least squares (KRLS) regression that allows for higher order terms and interactions without dimensionality problems (Hainmuller and Hazlett 2014). We collapse the data to classroom-level means and use the full set of factor-analysis mediators describe

above (the vector $\boldsymbol{M}'_{sco}$ in equations 3 to 5) to flexibly predict test scores. This estimator yields an R-squared of 0.80 for reading and 0.99 for writing (not shown), suggesting that the mediators are powerful predictors of test scores if we allow for a flexible functional form and interactions between them.

A potential concern with these estimates is overfitting: it is possible these R-squared values reflect strong predictive power within our sample that would not actually generalize to other datasets. The KRLS estimator is designed to mitigate overfitting by using leave-one-out cross-validation. If there are *K* observations it fits the model *K* times, in each instance leaving out one observation and computing the error in predicting the outcome for that observation. It then selects the functional form that minimizes the sum of the squared leave-one-out errors; the method thus provides a high degree of out-of-sample fit. Overfitting can still occur, however; in small samples (N<100), Hainmuller and Hazlett show that their estimated R-squared may be biased upward.

As a check on the potential for overfitting, we apply the estimator to random noise. If the estimator yields low R-squared values when applied to noise, then we can infer that it is finding real predictive power in our mediators. To do this test, we replace the real mediators with random numbers, using the same number of random variables as we have mediators in the real data (26 variables).[35] We then use the random numbers as "mediators" to see how well KRLS can use them to predict the outcome; we repeat the process 1000 times and examine the median R-squared. Replacing the actual data with random noise yields median R-squared values of 0.031 for reading and 0.031 for writing, suggesting that any upward bias in our estimates of the predictive power of the mediators is minimal.[36]

The machine-learning results reveal that the mediators have strong predictive power for

[35] Specifically, we draw 26 i.i.d. random variables from a U(0,1) distribution and apply kernel regularized least-squares to a model with the random variables on the right-hand side and test scores on the left. The R-squared values from these regressions give us a sense of the potential upward bias in the estimated R-squared for the real mediators.

[36] An alternative way to assess the empirical importance of overfitting is to use split-sample cross-validation. We randomly split the sample in half, estimating the model on one-half and assessing the fit (as measured by the R-squared) on the other. The drawback is that this requires estimating the model using a very small sample. We only have 72 classrooms with data on all the mediators, so a 50% random test sample contains just 36 observations. This is likely to be problematic for accurately assessing model fit: Harrell (2015) recommends that test samples have at least 100 observations. We repeat the split-sample approach 1000 times and focus on the median values of the distribution of estimates. We find that running the KRLS estimator on half of our sample gives a median R-squared of 0.78 for reading and 0.94 for writing. Constructing predicted values from the KRLS estimates and using them to predict the actual outcomes gives a median R-squared of 0.14 for reading and 0.25 for writing. The split-sample results suggest that KRLS could be over-fitting in the full sample, but the small sample sizes involved (just 36 observations) could be driving the problems we see there.

test scores if we allow for higher-order terms and interactions between the variables. Our tests suggest that this is unlikely to be the result of overfitting: applying the same estimator but exchanging the actual mediators for random noise yields close to zero predictive power. These results provide suggestive evidence that complementarities between inputs may be an important factor in understanding our results.

6.5 Discussion

A compelling explanation for our key results – large gains in the full-cost treatment group, and much smaller and even negative gains in the reduced-cost group – is that there may be strong complementarities between inputs in the teaching and learning process for reading and writing. These complementarities would mean that removing specific inputs could drastically change the returns in certain domains of learning (see for example, Mbiti et al. 2017). Although our experimental design does not allow us to conclusively estimate the importance of complementary inputs, our results are consistent with a model in which there are strong complementarities in the production of education.

Recall that the reduced-cost version of the program did not provide slates to schools. Data from the classroom observations show that students in the full-cost version were significantly less likely to spend time in class copying text from the board, and were more likely to be practicing writing on their own. In contrast, reduced-cost version students were more likely to practice writing using their hands "in the air." Taken together, our results on writing outcomes suggest strong complementarities between material inputs (slates) and worker human capital (teacher training and support).

At the same time, it is necessary to consider complementarities across skills to explain our results. Slates are most useful for the simplest writing tasks: writing letters, names, or single words. They are not ideal for writing entire sentences, let alone paragraphs. However, it is only for these advanced writing skills that the reduced-cost schools showed declines relative to the control group; simple writing skills measured by name-writing improved substantially in the reduced-cost study arm. If slates are indeed an important complement to other inputs in the NULP for the development of basic writing skills, students in the reduced-cost schools might take longer than those in the control to master the basics of writing letters and words. Basic writing is itself an important complementary input into the development of advanced writing skills. As a result, students without

slates would be unable to master writing sentences or paragraphs.

Separate evidence that the NULP's effects depend on a complex set of complementary inputs comes from our mediation analysis. Despite the significant differences between full-cost and reduced-cost schools in the classroom observations data, we can explain less than 5% of the difference in effects across program versions using those mediators linearly. However, using the kernel regularized least squares machine-learning estimator of Hainmueller and Hazlett (2014), the mediators have very strong predictive power for test scores, with R-squared values approaching one for writing. That estimator allows for higher-order terms and interactions – that is, for complementarities in the production function. It is likely that the correct functional form for estimating the role of the mediators does include such complementarities.

While the reduced-cost NULP's negative effects on writing are striking, the magnitude of the decline in reading impacts – compared to the effects of the full-cost NULP program – is nearly as large. Thus, complementarities between slates and the other NULP inputs are unlikely to be the entire story for our results. Instead, other differences – some of which would not seem qualitatively important ex-ante – are likely to be driving the differences in effectiveness across study arms. One key programmatic difference between the two variants of the NULP is the amount of follow-up that teachers receive each term. Teachers in the full-cost program schools receive over twice as many classroom support visits, which may help to reinforce teaching practices learned during training (Banerjee, Banerji, Berry, et al. 2016; Bruns and Luque 2015). Support visits can also serve as a form of informal monitoring; this type of incentive-free check-in with teachers has been shown to be effective in other education programs in Africa (Aker and Ksoll 2015).

Because we did not explicitly randomize each possible input combination in our study, we are unable to causally identify complementarities through reduced form analysis. Instead, our results provide evidence that is suggestive of, and consistent with, a complex and multi-dimensional learning process and the important role for complementarities in education production.

7 Conclusion

In this paper, we compare two versions of a primary literacy program, randomly assigned to schools in northern Uganda: a full-cost version delivered by the organization that designed the program, and a reduced-cost version delivered through a train-the-trainers approach, with some of

the more-expensive inputs removed. After one year, the full-cost version of the program leads to massive gains in learning. Reading improves by 0.64SD and writing by 0.45SD; these are among the largest learning gains ever measured for an education intervention. We see gains around 1SD for the most basic literacy skills: letter recognition and writing one's name. The reduced-cost version of the program fares much worse. It improves only basic reading and writing outcomes, leaving advanced reading skills unchanged and worsening students' advanced writing skills relative to the control group.

These qualitatively different outcomes arise from seemingly-minor differences in implementation and measurement details. Objective comparisons of the two program variants confirm that the differences are small: they differ by only 6% on a standardized metric of the attributes of in-service teacher-training programs (Arancibia, Popova, and Evans 2016). Yet, students in the reduced-cost version of the program experienced reading gains that were 80% smaller, and writing gains that were 135% smaller (that is, negative). Using detailed classroom observation data, we show that many aspects of time allocation, classroom management, pedagogy, and the use of materials vary significantly across treatment arms. However, when added linearly in mediation analysis, these variables do little to explain the differences in treatment effects. Less than 5% of the performance gap between program variants can be explained using our classroom observation mediators. If we instead use a machine-learning approach that allows for nonlinearities and interactions between mediators, we can explain an extremely large fraction of the variance in test scores – over 80% for reading and nearly 100% for writing. These results suggest that one compelling potential explanation for the sensitivity of the NULP's effectiveness is the presence of powerful complementarities in the education production function between inputs and across skills.

There has been surprisingly little research documenting complementarities in education. While non-experimental studies suggest that the estimated effectiveness of educational inputs is highly sensitive to functional form misspecifications (Figlio 1999), experimental evidence on complementarities is limited. Behrman et al. (2015) and Mbiti et al. (2017) find evidence in favor of complementarities while List, Livingston, and Neckermann (2013) do not. Few randomized evaluations have been designed to detect complementarities, let alone isolate their importance: the McEwan (2015) meta-analysis of education experiments in developing countries finds that only 9% of studies have more than two treatment arms, making it impossible to study complementarities

between inputs.

If there truly are powerful complementarities in the production of education, this may help explain the limited benefits of most education programs that have been evaluated by researchers. Three meta-analyses of hundreds of studies find average effects on test scores of less than 0.2SD, even when focusing just on the most-effective categories of interventions (Krisharatne, White, and Carpenter 2013, Conn 2017, and McEwan 2015). These are relatively small gains, especially given the low base of learning in developing countries. However, most studies evaluate just one educational intervention element – such as teacher training or textbook provision – and not programs that provide a package of interventions. If there are positive complementarities between inputs, the literature could be systematically underestimating the returns to specific educational investments. Suggestive evidence for this claim comes from the McEwan (2015) meta-analytic data: treatments with just one type of input (i.e., materials only, or teacher training only) have an average effect of 0.05SD, while those with more than one type of input (i.e., both materials and teacher training) have an average effect of 0.14SD. This difference is statistically significant, with a p-value of 0.008 (clustered by study).

Importantly, the complementarities suggested by our results are not necessarily complementarities in the underlying production function but combine both the structural parameters of the production function and agents' optimizing responses to variations in inputs (Glewwe et al. 2004). However, the existence of substantial reduced-form complementarities is still critical for policy choices, implying that program effectiveness is highly sensitive to small variations in certain inputs. This finding contributes to an ongoing debate about the validity of drawing inferences from experiments in economics and generalizability in randomized controlled trials. An extensive literature has criticized randomized experiments as being limited in their ability to guide policy and provide generalizable insights.[37] A growing body of research also documents that the effectiveness of social programs can be extremely sensitive to small differences in implementation, context, or measurement. Duflo (2017) argues that in a wide range of contexts, program details that do not seem economically important can matter a huge amount for the

[37] For instance, see Deaton (2010), Alcott (2015), and Banerjee et al. (2017) on threats to external validity, Ludwig, Kling, and Mullainathan (2011) on the difficulty of identifying mechanisms in most experiments, McEwan (2015) for a discussion of the lack of information about intervention costs, and Harrison and List (2004) and Levitt and List (2007) on the relative validity of lab and field experiments. Davis et al. (2017) discuss a potential solution to the problem of using a randomized experiment to study the how effective a program will be when implemented at scale.

practical results of a program. Vivalt (2017) finds substantial heterogeneity in the effectiveness of development programs that is robust across intervention type and specification. Taken together, these findings suggests that the body of evidence on "what works" using randomized trials lacks construct validity (Nadel and Pritchett 2016). This is a deeper issue than external validity: even if a program works equally well outside of the study setting, we may not be studying the same underlying object that would be implemented elsewhere.

Evidence on the sensitivity of program results to implementation details is scarce. A study by Bold et al. (2013) compares the effects of a contract teaching program implemented by an NGO with the same program implemented by the government; they find that the program significantly increases student test scores (by 0.18 standard deviations) when implemented by the NGO whereas the government version has no effect.[38] Similarly, Vivalt (2017) finds that government-implemented programs produce significantly lower effects. Our results verify and extend these previous findings: we show that changes to the details of a program that were quantitatively small using objective indicators (Arancibia, Popova, and Evans 2016), can not only drastically reduce its effectiveness, but actually cause *negative* impacts in certain areas. Moreover, our study is able to shed light on *why* different versions of the program have such different results. In the Bold et al. study, the different modes of program delivery are essentially "black boxes": we do not know what happened in the government-implemented vs. NGO-implemented versions that resulted in the difference in effectiveness. We use detailed classroom observations to examine how teacher and student behavior differs across study arms and use mediation analysis and machine learning to measure the importance of these differences for explaining test score gains (or losses).

Lastly, this study highlights the challenges of measurement in studying education programs. Metrics of learning vary widely across studies, and results are often compared in terms of standard deviations. Yet had we not measured both reading *and* writing outcomes, we would not have had a full picture of the effectiveness of the two versions of the program. Even within reading and writing, there is substantial variation in program impacts on different test components identifying the exact type, and level of difficulty of a skill. Researchers (especially economists)

[38] Another study showing the sensitivity of program effectiveness to program details in an entirely different context is Dhaliwal and Hanna (2017), which focuses on the health sector in India. An intervention to increase health worker attendance was effective only for nurses, and raised actual biometric health outcomes – but made doctors sufficiently unhappy that it was abandoned shortly after the successful randomized evaluation, because policymakers feared that in the long-run doctors would quit entirely.

should pay more attention to the type and administration of learning assessments.

Complementarities between inputs and the variety of outcomes available create a significant challenge for assessing the mechanisms behind a specific program's impacts. This also makes it considerably difficult for policy-makers and implementers to use research to inform educational policy and programs. Our experiment provides an extreme example, where the impact of a program is highly sensitive to small changes in inputs or outcome measures. Thus even successful educational interventions implemented as a pilot may be completely uninformative to the results of a scaled-up version of the program: it is hard to know whether a seemingly small change can cause a large difference in a program's impacts, and there are innumerable such changes that can and will occur.

A more-optimistic way of interpreting these findings is to focus on the fact that the full-cost NULP program produced *enormous* increases on students' reading and writing in grade one, after just a single year. This provides some hope that it is possible to produce substantial learning gains in the most poor, rural African schools, utilizing existing government teachers, and without offering monetary incentives or increases in wages.[39] Our findings suggest a direction for research on how to achieve gains like these. Programs designed to exploit complementarities – rather than to isolate the effects of individual inputs – are likely to be more effective at improving learning.

[39] This is in contrast to other programs that either recruit new teachers (Bold et al. 2013, Muralidharan and Sundararaman 2013, Duflo, Dupas, and Kremer 2015) or provide teachers with additional classroom help (Banerjee et al. 2007).

References

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *The American Political Science Review*, 110(3), 512.

Aker, J. C., & Ksoll, C. (2015). *Call Me Educated: Evidence from a Mobile Monitoring Experiment in Niger* (Working Paper No. 406). Center for Global Development.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165. https://doi.org/10.1093/qje/qjv015

Altinyelken, H. K. (2010). Curriculum change in Uganda: Teacher perspectives on the new thematic curriculum. *International Journal of Educational Development*, *30*(2), 151–161.

Arancibia, V., Popova, A., & Evans, D. K. (2016). *Training Teachers on the Job: What Works and How to Measure it* (Working Paper No. ID 2848447). Washington, DC: World Bank.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, 131(3), 1415–1453. https://doi.org/10.1093/qje/qjw016

Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments*, 1, 73–140. https://doi.org/10.1016/bs.hefe.2016.10.003

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M, & Walton, M. (2016). *Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India* (Working Paper No. 22746). National Bureau of Economic Research. https://doi.org/10.3386/w22746

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, 31(4), 73–102. https://doi.org/10.1257/jep.31.4.73

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy*, 123(2), 325–364. https://doi.org/10.1086/675910

Black, D. A., & Smith, J. A. (2006). Estimating the returns to college quality with multiple proxies for quality. *Journal of Labor Economics*, *24*(3), 701–728.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education* (SSRN Scholarly Paper No. ID 2241240). Rochester, NY: Social Science Research Network.

Boone, P., Fazzio, I., Jandhyala, K., Jayanty, C., Jayanty, G., Johnson, S., Ramachandrin, V., Silva, F., & Zhan, Z. (2013). *The Surprisingly Dire Situation of Children's Education in Rural*

West Africa: Results from the CREO Study in Guinea-Bissau (Comprehensive Review of Education Outcomes) (No. w18971). National Bureau of Economic Research.

Bruhn, M., & McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics, 1*(4), 200–232.

Bruns, B., & Luque, J. (2015). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: The World Bank.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics, 90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

Carrillo, P. E., Onofa, M., & Ponce, J. (2010). *Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador* (Working Paper No. IDB-WP-223).

Chao, M. M., Dehejia, R. H., Mukhopadhyay, A., & Visaria, S. (2015). *Unintended Negative Consequences of Rewards for Student Attendance: Results from a Field Experiment in Indian Classrooms* (SSRN Scholarly Paper No. ID 2597814). Rochester, NY: Social Science Research Network.

Conn, K. M. (2017). Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research*, in press. https://doi.org/10.3102/0034654317712025

Davis, J. M. V., Guryan, J., Hallberg, K., & Ludwig, J. (2017). *The Economics of Scale-Up* (Working Paper No. 23925). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w23925

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature, 48*(2), 424–455. https://doi.org/10.1257/jel.48.2.424

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2011). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education. In P. W. Glewwe (Ed.), *Education Policy in Developing Countries*.

Dhaliwal, I., & Hanna, R. (2017). The devil is in the details: The successes and limitations of bureaucratic reform in India. *Journal of Development Economics, 124*, 1–21. https://doi.org/10.1016/j.jdeveco.2016.08.008

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*. https://doi.org/10.1016/j.ijedudev.2014.11.004

Duflo, E. (2017). Richard T. Ely Lecture: The Economist as Plumber. *American Economic Review, 107*(5), 1–26. https://doi.org/10.1257/aer.p20171153

Evans, D. K., & Popova, A. (2016a). Cost-Effectiveness Analysis in Development: Accounting for Local Costs and Noisy Impacts. *World Development*, 77, 262–276. https://doi.org/10.1016/j.worlddev.2015.08.020

Evans, D. K., & Popova, A. (2016b). What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *The World Bank Research Observer*, *31*(2), 242–270. https://doi.org/10.1093/wbro/lkw004

Figlio, D. N. (1999). Functional Form and the Estimated Effects of School Resources. *Economics of Education Review*, *18*(2), 241–252.

Fryer Jr., R. G., & Holden, R. T. (2012). *Multitasking, Learning, and Incentives: A Cautionary Tale* (Working Paper No. 17752). National Bureau of Economic Research.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics*, 74(1), 251–268.

Glewwe, P., & Muralidharan, K. (2015). *Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications* (Working Paper No. 15/001). Research on Improving Systems of Education (RISE).

Glewwe, P., Ross, P. H., & Wydick, B. (2016). *Developing Hope: The Impact of International Child Sponsorship on Self-Esteem and Aspirations* (Working Paper).

Hainmueller, J., & Hazlett, C. (2014). Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Analysis*, *22*(2), 143–168. https://doi.org/10.1093/pan/mpt019

Harrell Jr., F. E. (2015). Model Validation. In *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis* (pp. 109–116). Switzerland: Springer.

Harrison, G. W., & List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, *42*(4), 1009–1055. https://doi.org/10.1257/0022051043004577

Heß, S. (2017). Randomization inference with Stata: A guide and software. *The Stata Journal*, 17(3).

Hornberger, N. H., & Chick, J. K. (2001). Co-Constructing School Safetime: Safetalk Practices in Peruvian and South African Classrooms. In M. Heller & M. Martin-Jones (Eds.), *Voices of Authority: Education and Linguistic Difference* (pp. 31–55). Westport, CT: Greenwood Publishing Group.

JPAL. (2014). Student Learning | The Abdul Latif Jameel Poverty Action Lab. Retrieved May 12, 2015, from https://www.povertyactionlab.org/node/7

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187–200.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* (Policy and Practice Brief). Bill and Melinda Gates Foundation.

Kayabwe, S., Nabacwa, R., Eilor, J., & Mugeni, R. W. (2014). *The Use and Usefulness of School Grants: Lessons from Uganda* (IIEP Country Notes). Paris, France: International Institute for Educational Planning.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119. https://doi.org/10.1111/j.1468-0262.2007.00733.x

Kremer, M., Brannen, C., & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. *Science*, *340*(6130), 297–300.

Krishnaratne, S., White, H., & Carpenter, E. (2013). *Quality Education for All Children? What Works in Education in Developing Countries* (Working Paper No. 20). New Delhi: International Initiative for Impact Evaluation (3ie).

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *The Journal of Economic Perspectives*, *21*(2), 153–174.

List, J. A., Livingston, J. A., & Neckermann, S. (2013). *Harnessing Complementarities in the Education Production Function* (Working Paper).

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism Experiments and Policy Evaluations. *The Journal of Economic Perspectives*, *25*(3), 17–38. https://doi.org/10.1257/jep.25.3.17

Mbiti, I., Muralidharan, K., Schipper, Y., Manda, C., & Rajani, R. (2017). *Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania* (Working Paper). National Bureau of Economic Research.

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, *85*(3), 353–394.

Murnane, R. J., & Ganimian, A. J. (2014). *Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations*. Harvard University OpenScholar.

Nadel, S., & Pritchett, L. (2016). *Searching for the Devil in the Details: Learning About Development Program Design* (Working Paper No. 434). Center for Global Development.

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

Piper, B., Zuilkowski, S. S., & Ong'ele, S. (2016). Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya. *Comparative Education Review*, 000–000. https://doi.org/10.1086/688493

Pritchett, L. (2013). *The Rebirth of Education: Schooling Ain't Learning*. Washington, DC: Center for Global Development.

Pritchett, L., & Beatty, A. (2015). Slow Down, You're Going Too Fast: Matching Curricula to Student Skill Levels. *International Journal of Educational Development*, 40, 276–288. https://doi.org/10.1016/j.ijedudev.2014.11.013

Rossell, C. H., & Baker, K. (1996). The Educational Effectiveness of Bilingual Education. *Research in the Teaching of English*, 30(1), 7–74.

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

Ssentanda, M. E. (2014). The Challenges of Teaching Reading in Uganda: Curriculum Guidelines and Language Policy Viewed from the Classroom. *Apples: Journal of Applied Language Studies*, *8*(2), 1–22.

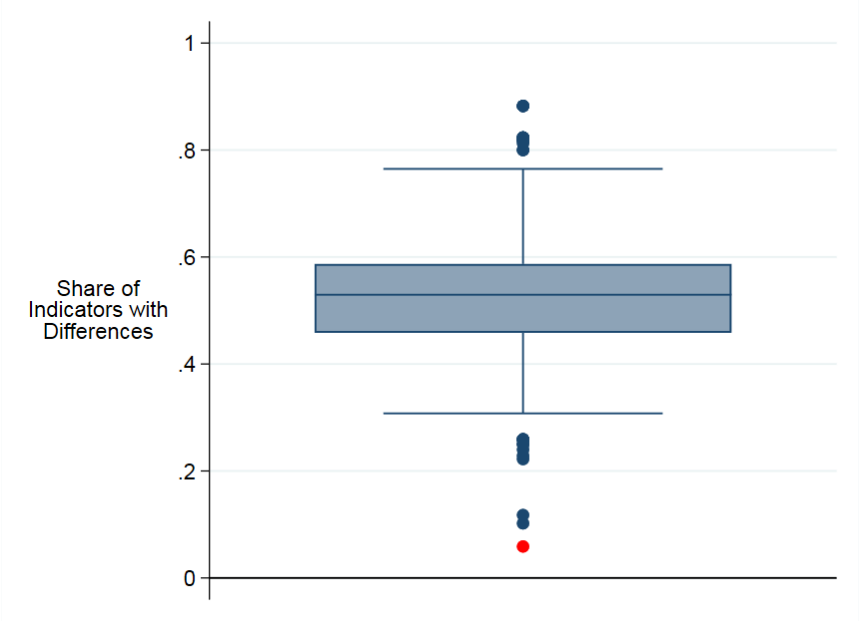
Tan, J.-P., Lane, J., & Lassibille, G. (1999). Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments. *The World Bank Economic Review*, *13*(3), 493–508.

Ugandan Ministry of Education and Sport. (2014). *Teacher Issues in Uganda: A Shared Vision for an Effective Teachers Policy*. UNESCO - IIEP Pôle de Dakar.

Vivalt, E. (2017). *How Much Can We Generalize from Impact Evaluations?* (Working Paper). Australian National University.

Webley, K. (2006). *Mother Tongue First: Children's Right to Learn in their Own Languages* (No. id21). Development Research Reporting Service, UK.

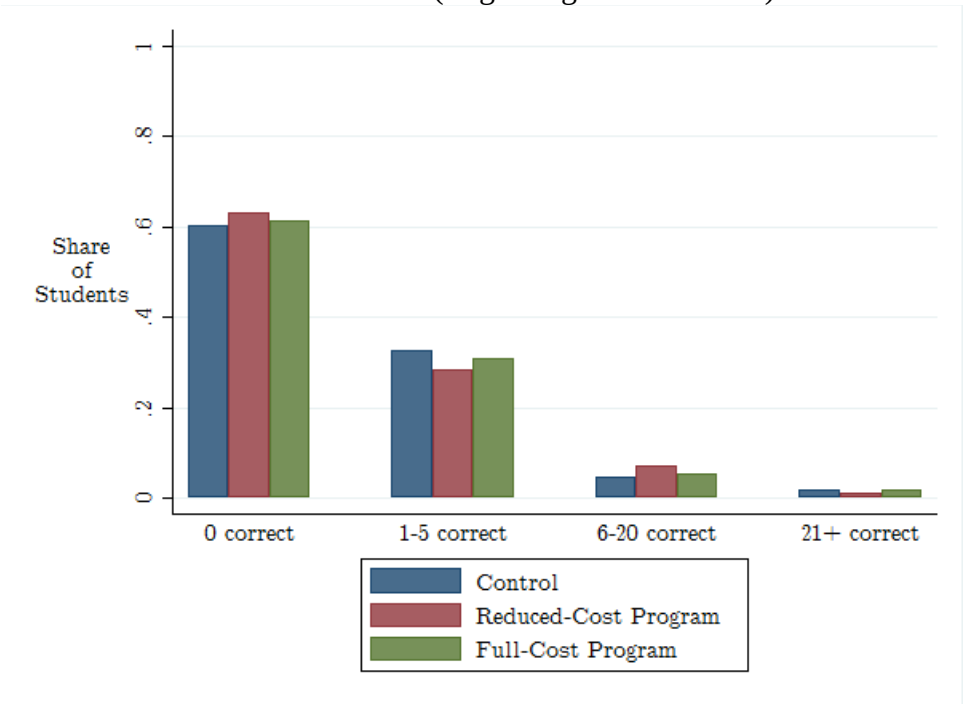
Figure 1: Box Plot of Differences across Programs in Arancibia, Popova, and Evans (2016)
Difference between NULP Variants Indicated in Red



Notes: This graph shows the share of indicators with any difference for all possible pairwise comparisons of programs in the Arancibia, Popova, and Evans (2016) data on the characteristics of in-service teacher training. There are 26 programs in their dataset and 325 possible pairwise comparisons (excluding comparisons of a program with itself). For each comparison, we compute how many of the 51 indicators of program characteristics have any difference across the two programs. (We exclude three indicators about sample size). The average pair of programs in the dataset differs on 53% of all indicators in the dataset. The two NULP variants differ on just 5.9% of indicators (shown in red above), which is the lowest value across all pairwise comparisons; the next-most-similar pair of programs differs on 10.2% of indicators.

Figure 2
Performance on Overall EGRA by Study Arm
(Number of Letters Correctly Recognized)

Panel A: Baseline (Beginning of First Grade)



Panel B: Endline (End of First Grade)

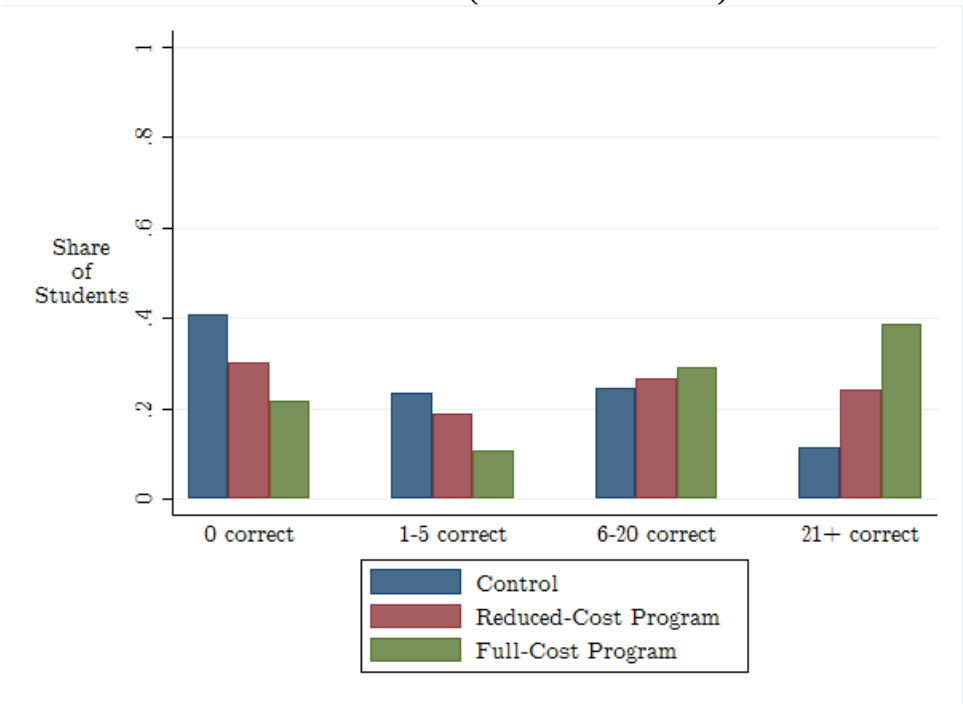
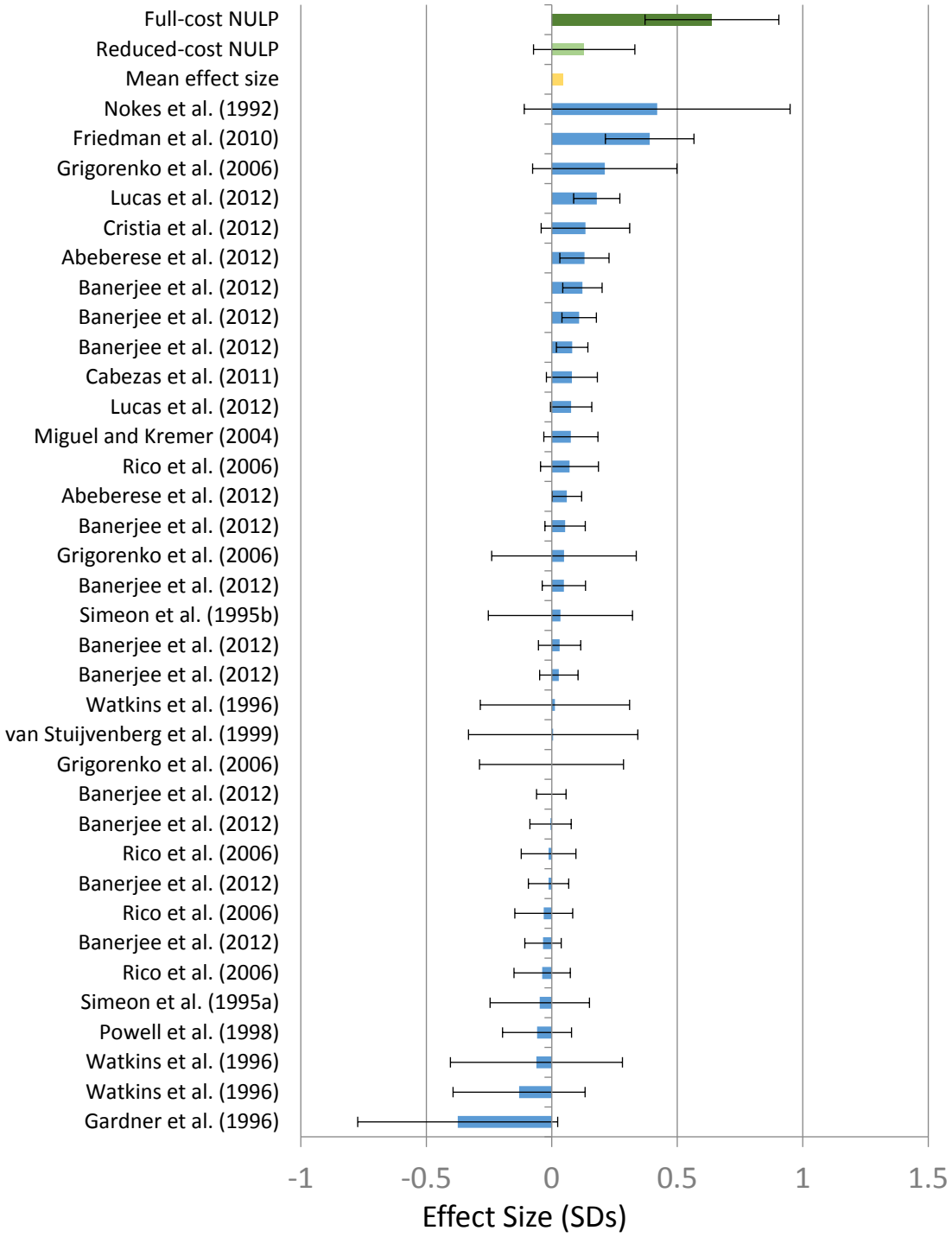
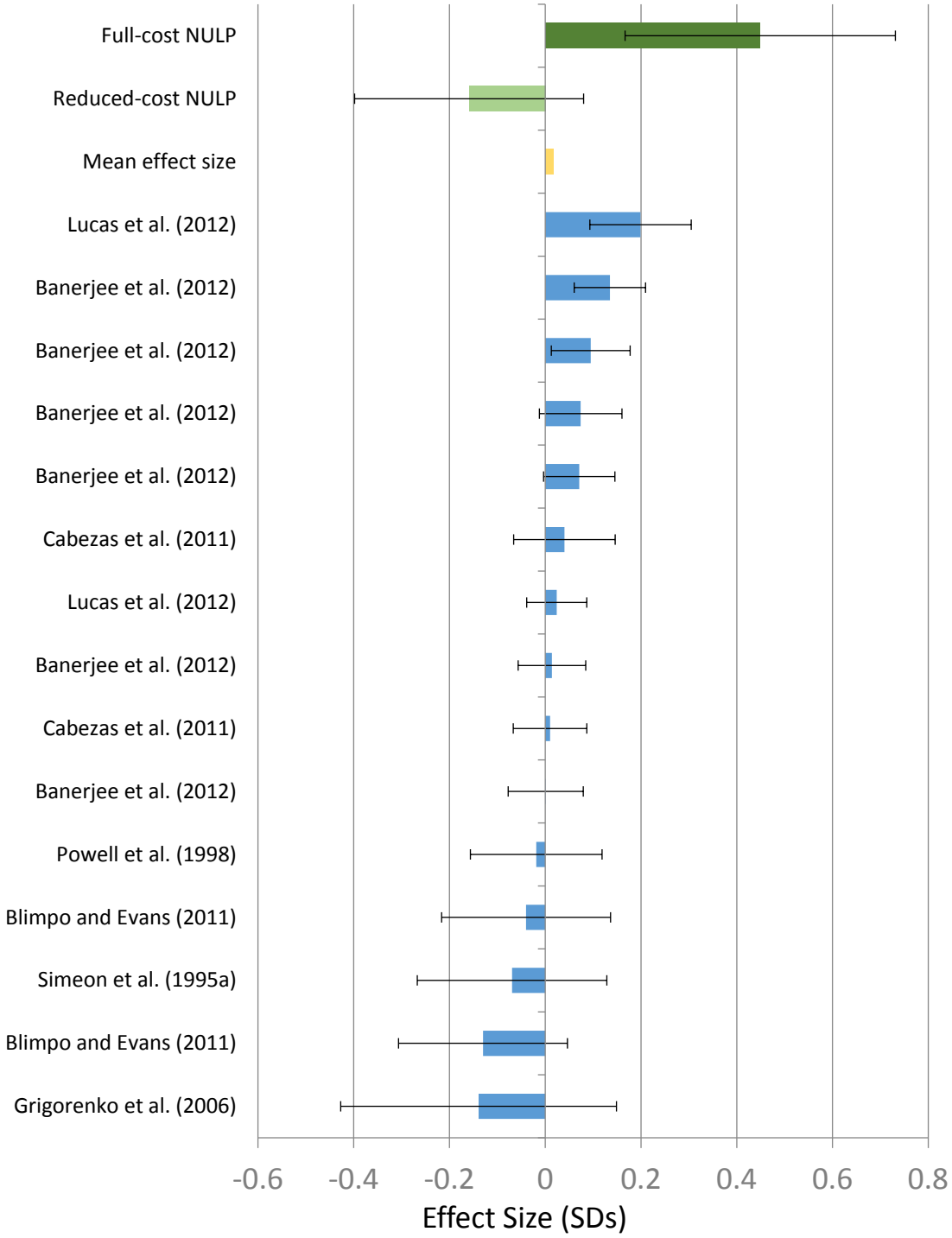


Figure 3
Comparison of Reading Impacts in McEwan (2015) Meta-Analysis



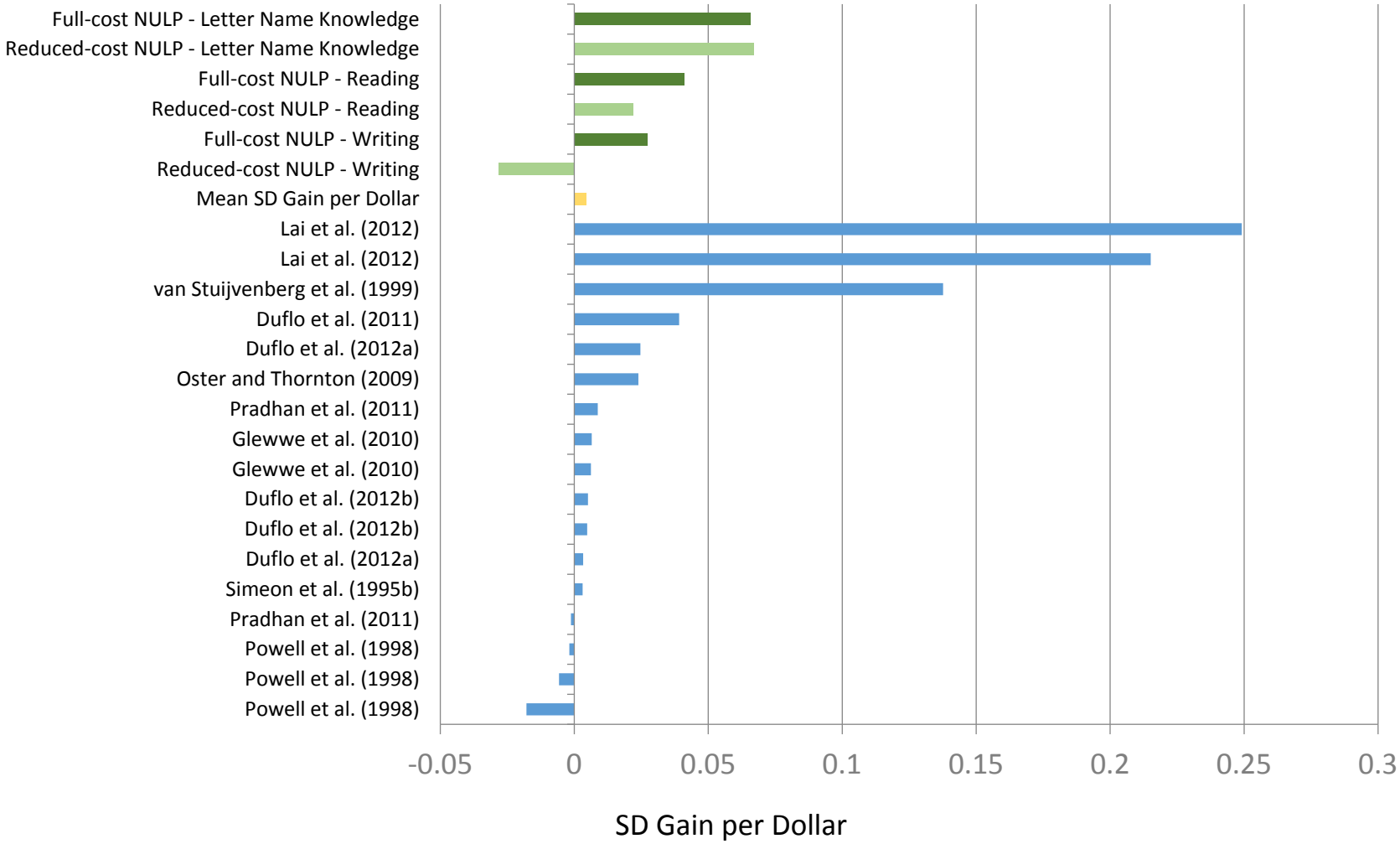
Notes: Graph of effect sizes for the two NULP variants as well as all programs from the McEwan (2015) meta-analysis where the outcome can be clearly identified as reading scores alone. Outcomes that appear to mix reading with other skills are not included. Repeated studies indicate that the same study included multiple experiments and/or outcomes. See the online appendix for a list of all the papers used in this figure.

Figure 4
Comparison of Writing Impacts in McEwan (2015) Meta-Analysis



Notes: Graph of effect sizes for the two NULP variants as well as all programs from the McEwan (2015) meta-analysis with an outcome that can be clearly identified as writing scores alone. Outcomes that appear to mix writing with other skills are not included. Repeated studies indicate that the same study included multiple experiments and/or outcomes. See the online appendix for a list of all the papers used in this figure.

Figure 5
Comparison of Cost-Effectiveness Estimates in McEwan (2015) Meta-Analysis



Notes: Graph of cost-effectiveness estimates (SD gains per dollar spent) for the two NULP variants as well as all programs from the McEwan (2015) meta-analysis with data on incremental costs. We exclude two of the three outcomes from Oster and Thornton (2009). Those outcomes have (insignificant) negative effects and extremely low incremental costs, leading to large negative cost-effectiveness estimates that make the figure difficult to read. See the online appendix for a list of all the papers used in this figure.

Table 1

Control Group Growth in Literacy During First Grade

	(1)	(2)	(3)	(4)	(5)
		Baseline		Change from Baseline to Endline	
	1(any correct)	Mean	SD	Mean	SD
Panel A: EGRA (Reading Test)					
PCA EGRA score index	0.394	0.000	0.808	0.148	1.034
Letter name knowledge (letters per minute)	0.153	1.180	4.424	4.857	9.349
Initial sound identification (sounds identified)	0.029	0.161	1.028	0.455	2.011
Familiar word reading (words per minute)	0.013	0.168	1.617	0.165	2.588
Invented word reading (words per minute)	0.006	0.084	1.191	0.275	2.309
Oral reading fluency (words per minute)	0.019	0.057	4.537	0.102	5.012
Reading comprehension (questions correct)	0.300	0.327	0.559	-0.111	0.703
Panel B: Writing Test					
PCA writing score index	0.237	0.010	0.161	0.468	0.902
African name (surname) writing	0.201	0.201	0.401	0.392	0.654
English name (given name) writing	0.145	0.145	0.352	0.193	0.499
Ideas	0.006	0.006	0.079	0.135	0.360
Organization	0.002	0.002	0.046	0.284	0.589
Voice	0.000	0.000	0.000	0.164	0.393
Word choice	0.069	0.069	0.254	0.099	0.374
Sentence fluency	0.006	0.006	0.079	0.261	0.584
Conventions	0.000	0.000	0.000	0.116	0.339

Notes: Statistics are for the 477 control-group members of the longitudinal sample, which includes students who were tested at baseline as well as endline. For the PCA indices, "any correct" indicates that the student had any correct answers on the entire exam. Change from Baseline to Endline is the student's endline score on the component minus his or her baseline score.

Table 2

Program Impacts on Leblango Early Grade Reading Assessment Scores
(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Leblango EGRA Score Index†	(2) Letter Name Knowledge	(3) Initial Sound Recognition	(4) Familiar Word Recognition	(5) Invented Word Recognition	(6) Oral Reading Fluency	(7) Reading Comprehension
Full-cost program	0.638***	1.014***	0.647***	0.374**	0.215	0.476**	0.445**
S.E.	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.128)	(0.113)
R.I. p-value	[0.005]	[0.006]	[0.007]	[0.010]	[0.161]	[0.025]	[0.030]
Reduced-cost program	0.129	0.407	0.076	-0.002	0.031	0.071	0.045
S.E.	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
R.I. p-value	[0.327]	[0.106]	[0.415]	[0.994]	[0.675]	[0.444]	[0.668]
Number of students	1460	1476	1481	1474	1471	1467	1481
Adjusted R-squared	0.149	0.219	0.103	0.066	0.075	0.074	0.058
Difference between treatment effects	0.509**	0.607**	0.570***	0.376***	0.184	0.405**	0.400**
S.E.	(0.127)	(0.159)	(0.128)	(0.092)	(0.093)	(0.117)	(0.120)
R.I. p-value	[0.010]	[0.020]	[0.006]	[0.007]	[0.212]	[0.021]	[0.038]
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

† PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Table 3

Program Impacts on Writing Test Scores

(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	Name-Writing		Story-Writing						
		African (Family) Name	English (Given) Name	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Full-cost program	0.449*	0.922***	1.312***	0.163	0.441	0.152	0.175	0.383	0.221	0.139
S.E.	(0.144)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
R.I. p-value	[0.064]	[0.001]	[0.001]	[0.536]	[0.173]	[0.539]	[0.466]	[0.231]	[0.385]	[0.558]
Reduced-cost program	-0.159	0.435**	0.450**	-0.274	-0.316	-0.313***	-0.262	-0.330	-0.253	-0.330***
S.E.	(0.122)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
R.I. p-value	[0.421]	[0.011]	[0.021]	[0.150]	[0.155]	[0.006]	[0.102]	[0.104]	[0.297]	[0.007]
Number of students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Adjusted R-squared	0.352	0.240	0.236	0.174	0.304	0.177	0.200	0.302	0.164	0.171
Difference between treatment effects	0.608***	0.487**	0.861***	0.436***	0.757***	0.465***	0.437***	0.713***	0.474***	0.469***
S.E.	(0.128)	(0.135)	(0.154)	(0.148)	(0.173)	(0.118)	(0.139)	(0.174)	(0.151)	(0.115)
R.I. p-value	[0.004]	[0.029]	[0.001]	[0.005]	[0.000]	[0.003]	[0.008]	[0.001]	[0.005]	[0.003]
Raw (unadjusted) values§										
Control group mean	0.482	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control group SD	1.000	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Presentation (column 10), which was not one of the marked categories at baseline; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Table 4

Cost-Effectiveness Calculations

	Program Variant	
	Full-cost	Reduced-cost
Cost per student per year	$15.39	$6.05
Letter Name Knowledge		
Effect size (SDs)	1.01	0.41
Cost per student/0.2 SDs	$3.04	$2.98
SDs per dollar	0.07	0.07
PCA EGRA Index		
Effect size (SDs)	0.63	0.13
Cost per student/0.2 SDs	$4.85	$9.10
SDs per dollar	0.04	0.02
PCA Writing Test Index		
Effect size (SDs)	0.42	-0.17
Cost per student/0.2 SDs	$7.29	N/A
SDs per dollar	0.03	-0.03

Notes: Costs based on authors calculations from actual expenditures by Mango Tree on each program variant in 2013. Only incremental costs are considered, and not costs related to materials development, curriculum design, etc. Effect size estimates come from our main analyses in Tables 2 and 3.

Table 5

Classroom Activities

	(1)	(2)	(3)	(4)
	Share of Time:			
	Reading	Writing	Speaking and Listening	Percent in Leblango
Full-cost program	0.065**	-0.036	-0.029	0.111*
S.E.	(0.014)	(0.017)	(0.013)	(0.036)
R.I. p-value	[0.013]	[0.148]	[0.102]	[0.054]
Reduced-cost program	0.054**	-0.003	-0.051**	0.079
S.E.	(0.014)	(0.016)	(0.014)	(0.039)
R.I. p-value	[0.017]	[0.902]	[0.020]	[0.207]
Number of observation periods	1288	1288	1288	1285
Adjusted R-squared	0.078	0.063	0.131	0.126
Difference between treatment effects	0.011	-0.033	0.022	0.032
S.E.	(0.015)	(0.016)	(0.012)	(0.029)
R.I. p-value	[0.644]	[0.215]	[0.193]	[0.398]
Control group mean	0.321	0.245	0.426	0.688
Control group SD	0.278	0.320	0.255	0.372

Notes: Sample is 1288 observation blocks, based on 440 individual lesson observations for 38 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

Table 6

Classroom Observations: Reading Elements and Materials

	(1)	(2)	(3)	(4)	(5)	(6)
	Element of Focus				Materials Used	
	Sounds	Letters	Words	Sentences	Primer	Reader
Full-cost program	0.115**	0.021	0.003	0.073	0.168**	0.062
S.E.	(0.026)	(0.034)	(0.026)	(0.047)	(0.039)	(0.030)
R.I. p-value	[0.013]	[0.691]	[0.946]	[0.321]	[0.011]	[0.305]
Reduced-cost program	0.073*	0.053	-0.014	0.002	0.110**	0.040
S.E.	(0.025)	(0.035)	(0.029)	(0.054)	(0.037)	(0.030)
R.I. p-value	[0.069]	[0.309]	[0.641]	[0.986]	[0.046]	[0.285]
Number of observation periods	893	893	893	893	893	893
Adjusted R-squared	0.054	0.025	0.045	0.044	0.091	0.222
Difference between treatment effects	0.042	-0.032	0.017	0.071	0.057	0.022
S.E.	(0.023)	(0.033)	(0.025)	(0.036)	(0.040)	(0.026)
R.I. p-value	[0.287]	[0.469]	[0.649]	[0.148]	[0.367]	[0.655]
Control group mean	0.064	0.217	0.839	0.442	0.024	0.052
Control group SD	0.246	0.413	0.368	0.498	0.154	0.223

Notes: Sample is 893 observation blocks in which students do any reading, based on 440 individual lesson observations for 38 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 7

Classroom Observations: Reading Class Indices from Factor Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Classroom Management					Pedagogy		
	Keeps Students Focused	Solid Lesson Plan	Active Throughout Classroom	Sounds and Letters Only	Whole Language On Board	Basic Elements in Breakout Sessions	Leblango Sentences in Reader	Paragraphs in Primer
Full-cost program	-0.138	0.061	0.049	0.096	-0.315**	-0.102	0.263**	0.161*
S.E.	(0.089)	(0.053)	(0.066)	(0.063)	(0.074)	(0.053)	(0.059)	(0.049)
R.I. p-value	[0.277]	[0.469]	[0.509]	[0.279]	[0.022]	[0.260]	[0.018]	[0.054]
Reduced-cost program	-0.160	0.018	-0.044	0.123	-0.076	-0.050	0.162*	0.121*
S.E.	(0.084)	(0.053)	(0.054)	(0.069)	(0.061)	(0.065)	(0.060)	(0.050)
R.I. p-value	[0.220]	[0.832]	[0.529]	[0.237]	[0.398]	[0.568]	[0.071]	[0.084]
Number of observation periods	890	890	890	893	893	893	893	893
Adjusted R-squared	0.091	0.128	0.219	0.043	0.111	0.187	0.152	0.146
Difference between treatment effects	0.023	0.042	0.093*	-0.027	-0.239*	-0.052	0.100	0.040
S.E.	(0.098)	(0.045)	(0.041)	(0.055)	(0.074)	(0.060)	(0.054)	(0.043)
R.I. p-value	[0.865]	[0.504]	[0.095]	[0.708]	[0.052]	[0.538]	[0.230]	[0.549]
Control group mean	0.185	0.084	-0.057	-0.091	0.170	0.015	-0.208	-0.092
Control group SD	0.590	0.504	0.565	0.634	0.550	0.574	0.543	0.452

Notes: Sample is 893 observation blocks in which students do any reading, based on 440 individual lesson observations for 38 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

Table 8

Classroom Observations: Writing Elements and Materials

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Element of Focus					Materials Used			Source Material	
	Pictures	Letters	Words	Sentences	Name	Air Writing	On Slate	On Paper	Copying Text from Board	Writing Own Text
Full-cost program	0.097	-0.103	0.031	-0.036	0.250**	-0.035	0.319***	-0.194***	-0.202**	0.327**
S.E.	(0.062)	(0.058)	(0.064)	(0.066)	(0.064)	(0.031)	(0.054)	(0.045)	(0.063)	(0.064)
R.I. p-value	[0.304]	[0.210]	[0.752]	[0.676]	[0.023]	[0.350]	[0.008]	[0.004]	[0.037]	[0.014]
Reduced-cost program	0.133	0.045	0.149	-0.142*	0.148**	0.058	0.027	-0.016	-0.097	0.070
S.E.	(0.061)	(0.056)	(0.065)	(0.051)	(0.043)	(0.030)	(0.044)	(0.034)	(0.060)	(0.053)
R.I. p-value	[0.104]	[0.623]	[0.126]	[0.072]	[0.011]	[0.217]	[0.568]	[0.699]	[0.244]	[0.327]
Number of observation periods	539	539	539	539	539	539	539	539	539	539
Adjusted R-squared	0.038	0.097	0.090	0.113	0.187	0.038	0.220	0.156	0.145	0.210
Difference between treatment effects	-0.036	-0.149	-0.118	0.106	0.102	-0.093**	0.292***	-0.178***	-0.105	0.257***
S.E.	(0.044)	(0.053)	(0.048)	(0.056)	(0.057)	(0.028)	(0.055)	(0.051)	(0.062)	(0.053)
R.I. p-value	[0.525]	[0.107]	[0.155]	[0.154]	[0.199]	[0.021]	[0.004]	[0.009]	[0.181]	[0.002]
Control group mean	0.329	0.348	0.594	0.290	0.174	0.148	0.052	0.806	0.658	0.265
Control group SD	0.471	0.478	0.493	0.455	0.381	0.357	0.222	0.396	0.476	0.443

Notes: Sample is 539 observation blocks in which students do any writing, based on 440 individual lesson observations for 38 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

Table 9

Classroom Observations: Writing Class Indices from Factor Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Classroom Management			Pedagogy				
	Keeps Students Focused	Solid Lesson Plan	Active Throughout Classroom	Pictures, Words, and Stories	Copying Teacher's Text	Leblango Practice on Slates	Pictures and Letters on Paper, High-Energy	Leblango Sentences and Handwriting
Full-cost program	-0.176	0.093	0.195**	0.196*	-0.473***	0.626***	-0.160*	-0.020
S.E.	(0.104)	(0.058)	(0.065)	(0.083)	(0.091)	(0.096)	(0.058)	(0.065)
R.I. p-value	[0.191]	[0.381]	[0.028]	[0.093]	[0.008]	[0.001]	[0.086]	[0.742]
Reduced-cost program	-0.165	0.129*	0.055	-0.036	-0.178	0.294***	0.034	-0.079
S.E.	(0.110)	(0.059)	(0.057)	(0.082)	(0.090)	(0.075)	(0.064)	(0.065)
R.I. p-value	[0.233]	[0.082]	[0.491]	[0.796]	[0.180]	[0.001]	[0.670]	[0.483]
Number of observation periods	537	537	537	539	539	539	539	539
Adjusted R-squared	0.086	0.250	0.188	0.107	0.213	0.294	0.087	0.227
Difference between treatment	-0.010	-0.037	0.140**	0.232*	-0.295**	0.332***	-0.194*	0.060
S.E.	(0.106)	(0.078)	(0.046)	(0.077)	(0.084)	(0.078)	(0.073)	(0.056)
R.I. p-value	[0.943]	[0.753]	[0.016]	[0.069]	[0.017]	[0.002]	[0.060]	[0.495]
Control group mean	0.086	-0.004	0.125	-0.128	0.191	-0.352	-0.017	-0.008
Control group SD	0.756	0.645	0.576	0.851	0.726	0.480	0.610	0.580

Notes: Sample is 539 observation blocks in which students do any writing, based on 440 individual lesson observations for 38 schools. Observation windows are typically 10 minutes long, but can vary in length if the class runs long or ends early. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

Table 10

Mediation Analysis

	(1)	(2)	(3)
	Letter Name Knowledge	PCA Leblango EGRA Score Index	PCA Writing Score Index
Demediated Treatment Effect			
Difference between full-cost and reduced-cost programs	0.681***	0.598***	0.645***
S.E.	(0.127)	(0.095)	(0.101)
R.I. p-value	[0.002]	[0.000]	[0.000]
Adjusted R-squared	0.232	0.159	0.331
Number of observations	15,516	15,311	14,559
Share of treatment effect explained by mediators	0.011	0.020	0.037
Raw (unadjusted) values[§]			
Reduced-cost program mean	11.346	0.31	-0.054
Reduced-cost program SD	13.861	1.072	0.639

Notes: Sample is the combination of each student with all classroom observation windows for that student's class; re-estimating our main regressions on this modified sample yields similar treatment effects and confidence intervals to the main sample. The analyses in this table are restricted to data from the two treatment arms. We estimate the demediated treatment effect using the sequential *g* estimator of Acharya et al. (2016), by removing the effect of the treatment on the mediators from the outcome and then regressing the demediated outcome on the treatment indicator. Reduced-Cost Program Mean and SD are computed using the endline data for the reduced-cost group alone. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.