Making the Grade:
The Trade-off between Efficiency and Effectiveness in Improving Student Learning

Jason T. Kerwin and Rebecca L. Thornton*

July 14, 2017

Click here for the latest version of this paper

Abstract

Relatively small changes to the inputs used in education programs can drastically change their effectiveness if there are large trade-offs between effectiveness and efficiency in the production of education. We study these trade-offs using an experimental evaluation of a literacy program in Uganda that provides teachers with professional development, classroom materials, and support. When implemented as designed, and at full cost, the program improves reading by 0.64 SDs and writing by 0.45 SDs. An adapted program with reduced costs instead yields statistically-insignificant effects on reading – and large *negative* effects on writing. Detailed classroom observations provide some evidence on the mechanisms driving the results, but mediation analyses show that teacher and student behavior can account for only 6 percent of the differences in effectiveness. Machine-learning results suggest that the education production function involves important nonlinearities and complementarities – which could make education programs highly sensitive to small input changes. Given the sensitivity of treatment effects to small changes in inputs, the literature on education interventions – which focuses overwhelmingly on stripped-down programs and individual inputs – could systematically underestimate the total gains from investing in schools.

* 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 University, Université Paris-Dauphine, the University of Minnesota, Centre for the Study of African Economies, Wilfrid Laurier University, the Comparative and International Education Society Annual Meeting in Vancouver, 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 appendices 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. A smaller but growing literature also examines the cost-effectiveness of various interventions.[1]

Despite the large body of evidence on "what works," there is little concrete guidance on how best to use these results to improve education systems. Even if policymakers were armed with a comprehensive list of proven programs, including their effect sizes and costs, they would still face trade-offs between program effectiveness, budget constraints, and equity. Limited resources make it infeasible to provide the highest quality – and presumably most effective – programs to all. To deliver quality universal education, is it better to implement the most cost-effective program so as to reach the largest number of students, or, to deliver the highest-impact programs, regardless of cost, to a limited population? A common approach is to pick an effective program and make it cheaper by modifying some of the most expensive inputs. This option is often tempting, since effective interventions combine numerous inputs, many of which may seem unimportant or overly costly. However, this strategy could lead to qualitative differences in program impacts if there is a large trade-off between effectiveness and efficiency in the production of education – for example, if there are important complementarities between inputs.

In this paper we examine the trade-off between effectiveness and efficiency for a primary literacy program in Northern Uganda. To do this, we compare two versions of the program: the first version is administered at full cost, with the program's entire array high-quality inputs, while

[1] Evans and Popova (2016b) discuss six systematic reviews of the effectiveness of education programs in developing countries: Conn (2017); Glewwe et al. (2014); Kremer, Brannen, and Glennerster (2013); Krishnaratne, White, & Carpenter (2013); McEwan (2015); and Murnane & Ganimian (2014). Since Evans and Popova (2016b) conducted their literature review, at least one additional review has been released (Glewwe and Muralidharan 2015). The literature on cost-effectiveness is far sparser, largely because of the limited availability of data on program costs. Kremer, Brannen, and Glennerster (2013), McEwan (2015) and Dhaliwal et al. (2011) do a systematic comparison of cost-effectiveness across studies; Glewwe and Muralidharan (2015) and Krishnaratne, White, and Carpenter (2013) both include some discussion of cost-effectiveness.

the second version is delivered with some modifications to the inputs at reduced cost, as the program would likely be delivered at scale. We experimentally estimate the effects of these two program variants on student learning by randomly assigning 38 government primary schools to one of three study arms – the full version of the literacy program, the reduced-cost version of the program, and a control group. We then use the estimated treatment effects of each variant and detailed data on their costs, to compare their cost-effectiveness.

The program we study, 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 early-grade primary school teachers, to teach children to read and write in their native language before transitioning 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 and in-service training, orthography workshops, mentor teacher support, and monthly classroom visits. The trainings and support are delivered by program staff, most of whom had been expert teachers, trainers, and government tutors prior to working for the implementing organization. This paper studies the impacts of the NULP at the end of grade 1 (P1). The literacy program – when delivered at full cost – is highly effective, even after just 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 (both name-writing and story-writing, as measured by a summary index) by 0.45 standard deviations. These reading and writing effects are comparable to some of the largest measured in the literature, with confidence intervals bounded well away from zero.[2] While highly effective, this full-cost model is very costly for a developing-country education intervention, at about $15 per student.[3] If the government wanted to scale-up the program to all students in the Lango sub-region, it would cost over one-quarter of that sub-region's total primary

[2] We can rule out effects smaller than 0.68 for letter-name knowledge, 0.37 SDs for overall reading, 1.03 for first-name writing, 0.78 for surname writing, and 0.17 for overall writing. These confidence intervals are both farther from zero and tighter than most of the literature: the average test score gain across the 71 teacher-training interventions covered in McEwan (2015) is just 0.12 SDs. Two of those interventions have effect sizes comparable to the impacts we estimate, although the estimates are quite noisy: Carrillo et al. (2010) cannot reject 0 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.

[3] The per-student cost is higher than 93% of the programs reviewed in McEwan (2015).

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. How would this impact the effectiveness of the program? To study this, the NULP was modified to reduce costs in a way that explicitly emulated how it might be delivered at scale. The changes included: 1) removing the most expensive material inputs (slates and clocks), and 2) using a cascade model of delivery resulting in fewer days of training and support visits. 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]

These small modifications generate qualitatively different conclusions about the effectiveness of the program. We find improvements in letter name knowledge in the reduced-cost version of the program, but they were considerably smaller than those in the full-cost program schools (0.41 standard deviations). Evaluated on the basis of letter name gains alone, the reduced-cost version of the NULP is slightly more cost-effective than the full-cost version. The cost-effectiveness results are very different when we look at more-sophisticated literacy skills, where we find no distinguishable gains to reading actual words or sentences in the reduced-cost program. Gains to overall reading scores are small and statistically insignificant (0.13 standard deviations, p =0.218).

The results of the original and modified programs diverge even further when we examine writing outcomes. The modified program shows gains only for the most basic writing tasks - 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: each component of the writing assessment drops by about 0.3 standard deviations, and the drop is significant at the 0.05 level for voice, word choice, and sentence fluency.[5]

Drawing on a rich set of classroom observations and using factor analysis methods, we document important differences in classroom management, pedagogy, and the use of materials

[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 shows that education interventions can have unanticipated negative consequences (Chao et al. 2015, Fryer and Holden 2012). Both of these examples involve providing explicit rewards for performance. In contrast, the NULP provides no extrinsic incentives to students or teachers.

across the three study arms. For example, teachers in the full-cost version of the program are more likely to move throughout the classroom to keep entire class engaged and less likely to do 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.

However, these large differences in pedagogy and classroom management do little to explain why the full-cost treatment was so much more effective. To examine the role of the differences in classroom behavior in driving the gap in treatment effects, we conduct mediation analyses using the Acharya et al. (2016) sequential *g*-estimator. We find that observable changes in behavior explain only 6% of the overall difference in effectiveness across the two program variants; this is true for both reading and writing outcomes. This is not because the classroom observation data explain test scores poorly: the adjusted R-squared from regressing classroom-average reading scores on the full set of mediators is 0.13, and the corresponding figure for writing scores is 0.28. A machine learning approach suggests that our measured mediators have even higher predictive power for test scores. Applying the kernel regularized least squares approach of Hainmueller and Hazlett (2014) yields an R-squared of 0.82 for reading and close to 1 for writing.

Our findings suggest that the effectiveness of an intervention can be highly sensitive to small changes in inputs and argue for caution when modifying programs to reduce costs, adapt them to new contexts, or go to scale.[6] 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 students *worse* off in some areas than they would have been under the status quo.

Implementing an effective program necessarily involves a large number of highly-contextual decisions about a wide range of inputs. In some cases, certain inputs are obvious complements to others, especially to practitioners who design curriculum and education programs. In the full NULP, for example, slates are given to P1 students to complement the teacher training in the process of teaching writing. The reduced-cost program did not provide slates, which is likely one contributor to its negative effects on writing. How various inputs interact with each other may

[6] In the language of Nadel and Pritchett (2016), the design space is "rugged". Our findings imply an even stronger result – that some aspects of the design are important for results, even though their role may not be immediately obvious. This parallels the point made by Duflo (2017), who argues that economists should focus on the "plumbing" of programs: details that seem economically unimportant but have major logistical implications, and can matter a great deal for a program's effectiveness.

not, however, be obvious based on *ex ante* reasoning or experience. The drop in effectiveness in the reduced-cost program is nearly as large for reading as it is for writing, for example, but no obvious complementary inputs were removed from the reading production process.

Given the intrinsically multi-dimensional, path-dependent nature of the learning process and the many potential complementarities between inputs, it may be infeasible to know which specific components of a program are most important for its success. To tease out the effects of various combinations of inputs would require a large number of experiments, and could necessitate variation in inputs that cannot or should not be altered. For example, teachers in the full-cost NULP received far more classroom support visits (where an expert observed their class and provided feedback) than their counterparts in the reduced-cost version. While cost was a factor in this decision, another was that the support visits were being provided by government employees with numerous other demands on their time. Perhaps the additional support visits were an important complement to the other inputs, helping teachers actually implement the curriculum as designed. The stronger the complementarity between two program inputs, the more difficulty it is to measure. The strongest complementarities cannot be studied at all: for example, the teacher training relies upon the textbooks in its curriculum, so there is no way to offer the training without the textbooks. We therefore cannot hope to measure the importance of the complementarity between the textbooks and the training for the program's results.

More broadly, the sensitivity of the program's effectiveness to small changes in implementation suggests that the current approach to evaluating education interventions may be flawed. Randomized evaluations have heretofore typically focused on trying 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". Also, evaluations are often conducted on versions of programs that have already eliminated costly inputs and made other design choices. Our findings imply that for both of these reasons, the literature is likely to systematically underestimate what can be achieved by investing in education in the developing world. They also imply that the goal of providing policymakers with a list of pre-packaged cost-effective interventions may be a false one. The details matter immensely.

2 Context and Intervention

2.1 Primary Education in Uganda

The literacy program we study provides a bundle of high-quality educational inputs designed to address the needs of teachers and students in rural Ugandan government primary schools. Before turning to the details of the program, we describe the broader context of the Ugandan education system. Primary education in Uganda consists of seven years of schooling for children ages 6-12, with students taking a Primary Leaving Examination (PLE) at the end of grade 7. Grades P1-P3 (1st to 3rd grade) are designated "early primary" and grades P4-P7 (4th to 7th grade) as "upper primary"; each year of school consists of three terms of equal length. The program we evaluate focuses entirely on the early primary grades in government schools.

Since the government implemented Universal Primary Education (UPE) in 1997, primary school has been free. Schools receive funding from the government via the UPE capitation grant program, launched at the same time as the overall UPE program. These grants are intended in part to pay for instructional materials (e.g., books, chalk, wall charts, teachers' guides, and resource books). Often, however, the number, delivery, and use of material resources are inadequate.[7]

Teachers in Ugandan primary schools receive their basic teacher training via a certificate program available at primary teacher colleges across the country. Entering the certificate program requires prospective teachers to have finished their O-Level exams (lower secondary school) with a minimum number of passed courses (Uganda Ministry of Education, 2014). Once hired, teachers receive additional training and continuous professional development through the Teacher Development and Management System (TDMS). Under this system, Coordinating Centre Tutors (CCTs) conduct in-service trainings and provide support supervision of teachers: classroom visits that provide feedback and guidance on teaching best practices.[8]

The TDMS often works with a cascade – or "train-the-trainer" approach to training – intending that trainers pass on skills and competences to CCTs, who then directly train teachers.

[7] A survey of 14 schools across Uganda found report that all the schools "were critically in need of instruction materials, particularly textbooks" (Kayabwe, 2014). Severe shortages of materials are apparent in our study sample as well. Control-group classrooms engaged in writing had enough paper and pencils (or alternatively enough slates and chalk) for all the students just 50% of the time they were observed. Textbooks were even scarcer: control classrooms had enough textbooks for every student in writing class just 16% of the time, and 78% of the time had no textbooks at all.

[8] These employees are known as Coordinating Centre Tutors, or CCTs, because each one manages a set of schools near an administrative office known as a Coordinating Centre, or CC.

There has been little change in the content of the teacher education curriculum in the past decade. Many of the teaching methods rely on call and repeat, where the teacher will point to a sound, letter, or word on the board, say it, and students will then repeat. This call and repeat pattern can last for many minutes where the students tasks is to repeat and memorize whole words, rather than understand how to sound out words (Ssentanda, 2014).[9]

In 2007, the Ministry of Education approved a thematic curriculum in which children are taught content around familiar themes such as "our home" or "weather" (Altinyelken, 2010). The curriculum gives guidelines for literacy lessons, stipulating that children should have an hour each day to practice reading and writing.[10] Children in first grade are also taught half-hour classes of Oral Literature, News, and Oral English.

Since 1999, Uganda's official policy is that students in grades P1-P3 are to be taught in their local language with P4 used to transition to all English instruction. In practice – however, English is still heavily used as the de facto language of instruction across the country.[11] 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. 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 as a result, education systems that use a language unfamiliar to children in school, and simply hope that children will pick up that language, are failing. However, well-identified evidence about the causal effects of mother-tongue instruction is sparse.[12]

[9] Hornberger and Chick (2001) document similar classroom behavior patterns in Peru and South Africa. They argue that the extensive use of the call-and-repeat pattern is exacerbated by the extensive use of foreign languages in the schools they study: it allows for "participation" even when students do not understand, and reduces the risk that students' (and teachers') lack of understanding will be exposed.

[10] Literacy lessons include: Literacy I (which "focuses on reading, with presentations, practice, pre-reading activities and an emphasis on the sight words."), followed by Literacy II (which "focuses on prewriting activities, drawing, labelling and developing handwriting"). Teachers are also asked to spend the last 20 minutes of every literacy hour on writing, or "pattern practice."

[11] Uganda has 41 different languages from three different language families, many of which have poorly-developed orthographies and limited or nonexistent teaching and reading materials. These problems are compounded by deficient training of teachers in implementing mother-tongue instruction. As a result, teachers commonly revert to English as a language of instruction. In our data, P1 Classrooms in the control group conduct nearly 40% of reading and writing lessons in English rather than in the local language.

[12] A review by Rossell and Baker (1996) found that virtually all studies of the practice are of Spanish language instruction in the US – and no evidence from developing countries at all. In the US studies they review, only 22% find benefits for reading and there are rarely any benefits for other subjects. A recent randomized trial by Piper, Zuilkowski, and Ong'ele (2016) found that mother-tongue instruction in Kenya led to improved reading scores in one's own language by between 0.3 and 0.6 SD.

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 suffers severe infrastructure shortages, extreme poverty and poor access to quality education, especially after the area was devastated by civil war from 1987-2007. 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).

2.2 The Northern Uganda Literacy Project

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. The program was piloted from 2009 to 2012 and pedagogical, curricular, and logistical refinements were made to the model to improve its effectiveness. During this time, Mango Tree scaled up the program from a single pilot school to eight government primary schools.

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 addresses multiple issues at once, providing a carefully-designed bundle of complementary inputs. We first describe the program elements of the full program and the changes made to the reduced-cost version of the program. We then report how the two versions of program are coded according to the in-service teacher training tool developed by Arancibia, Popova, and Evans (2016).

The basis of the NULP model is mother tongue instruction, where children are taught first in the language that they grew up speaking. Mango Tree developed the literacy program 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 P1, entirely in the students' mother tongue. The NULP model introduces content slowly, providing time for repetition and revision. Sixteen of the twenty-five letters of the Leblango alphabet are taught in

P1, with the remainder taught in P2.[13] The slower pace of instruction in NULP classrooms could have important learning benefits: Pritchett and Beatty (2015) argue that more-ambitious curricula can actually hurt learning, because students are left behind and cannot catch back up. Teachers are provided with scripted lesson plans for each literacy lesson, providing teachers with easy-to-remember steps that become routine over time. Literacy lessons consist of: Word Building, Story Reading, Creative Writing, and English; Although oral English is given as a subject, students are not exposed to written English on the board or in reading materials.

Classrooms are provided a set of tailored materials including primers (textbooks that follow the curriculum and provide visual examples for students) and readers (books that provide text for reading practice) that are small, durable, and easy to store in the classroom. P1 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). Each classroom is provided with a wall clock to help teachers keep track of the time during a lesson.

The NULP provides extensive training and support for teachers in the program's classrooms throughout the school year using expert trainers, detailed facilitators' guides, and instructional videos. The first teacher training module is a five-day residential workshop, before the beginning of the school year, on the Leblango orthography, including grammatical features and letter names and sounds. Teachers also undergo three additional intensive, residential trainings on literacy methods before each of the three terms. In addition to the residential trainings, there are also six in-service training workshops on Saturdays throughout the school year.[14] The training sessions are integrated with the textbooks and teacher guides; trainers draw on them extensively during the training sessions. The training sessions are complemented by support supervision visits conducted three times each term that provide teachers with feedback about their teaching. The support supervision visits and training sessions are conducted by experienced MT staff members as well as "mentor teachers" with previous experience teaching the NULP instructional model. Teachers also receive two support supervision visits from CCTs each term. The CCTs are trained

[13] The letters taught in P1 are M, A, C, N, K, O, I, W, L, E, R, G, P, D, T, and Y. The letters taught in P2 are B, J, NY (which is a single letter in the Leblango alphabet), U, Ë, Ï, Ö, Ü, and Ŋ (which is sometimes written "NG", as in the word "Leblango".)

[14] This program engages with existing teachers, 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).

to provide the same type of feedback on teaching performance as the MT staff, so each teacher receives five total support supervision visits per term.

Two other aspects of the NULP model involve engagement outside of the classroom. First, 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.[15] Lastly, the NULP "Strengthening a Literate Society" program engages with communities more broadly by supporting the Language Board to standardize orthography, printing and distribute basic language reference materials, training local writers, editors, and illustrators to develop local language literacy materials, and airing a radio program that does local-language literacy promotion. 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.

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, 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% of the studies in the McEwan sample.[16] 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.[17]

2.3 The Reduced-Cost Version

Mango Tree's goal in developing the NULP over three years and beginning initially in just a handful of schools was to create the highest-quality literacy program possible, given the multiple challenges that teachers and students face 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 version of the NULP that was explicitly designed to resemble

[15] 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.

[16] These figures are based on only 16 programs; the overwhelming majority of studies in the McEwan sample do not provide incremental cost information.

[17] 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.

the way the program could be implemented at scale.

There are three main differences between the full-cost and modified versions of the NULP. The first modification is the use of a cascade model of training and support, rather than working directly with teachers. The cascade, or "train-the-trainers" model of delivery involved Mango Tree program 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. In general, CCTs are only mandated by the government to make classroom visits once a term and hold termly trainings consisting of one-day workshops.

Under the modified NULP model, 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. CCTs were given financial resources (for transportation and refreshments), to make school visits and to hold training sessions.

The second main difference between the full-cost and modified versions of the NULP is that schools in the reduced-cost version of the program received fewer support visits than those in the full cost version: two visits per term instead of five. Teachers also received training in their local communities, rather than off-site residential trainings in the district capital. Lastly, classrooms in the modified version were not provided with slates and wall clocks, which were seen as expensive and less-essential inputs for the program. The inputs provided to schools in each version of the program are listed in Appendix Table A1. In all, the modifications to the full program reduced the program's 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 the most important elements of in-service teacher training programs. Using written documentation and oral interviews, they use their instrument to code 26 in-service training programs, including the two versions of the NULP (Appendix Table A2). Several of these indicators are particularly important for describing the NULP model and characterizing the differences between the full and modified versions. First, other than the total number of schools treated, there are no overarching aspects of the program that differ across the two versions. There are also no differences in content between the two versions. The third category, delivery, correctly reflects the main differences in the programs: direct vs. cascade delivery, profile of those delivery the training, and the number of

support visits to teachers after the initial training. In total, three indicators (6.25 percent) differ across the two versions of the NULP.[18]

3 Research Design

3.1 Sample and Randomization

We evaluate the impact of the two program variants on student learning using data from a stratified randomized controlled trial. The study was conducted 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 these criteria.[19] In addition, head teachers were asked to assign the two best early primary teachers in the school to each P1 classroom. Prior to the assignment of schools to study arms, each head teacher signed a contract with Mango Tree outlining the guidelines for participation in the evaluation. These contracts had credibility: Mango Tree had used them in previous years in schools where it was piloting the NULP, and schools that did not adhere to the contracts lost Mango Tree support.

The 38 schools in the study were assigned to one of three study arms via public lottery: control schools, full-cost program schools, and modified program schools. The lottery was held at a stakeholder meeting in late December 2012, to provide Mango Tree enough time to train teachers 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 P1 enrollment, and distance to the Coordinating Centre headquarters. Representatives from each school within a stratification cell drew tokens indicating treatment

[18] We exclude three indicators describing sample size. If we added an indicator for provision of slates, 8.33 percent of the indicators would differ across the two versions of the program.

[19] The criteria were: having two P1 classrooms and teachers, having desks and lockable cabinets for each P1 class; having a student-to-teacher ratio of no more than 135 during the 2012 school year in grades P1 to P3; being located less than 20 km from the CC 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.

status from an urn.

After the second week of the 2013 academic year, enumerators collected student enrollment rosters from each school. We used the names on the rosters to generate an ordered list of 70 randomly-selected students, stratified by classroom and gender. The first 50 students on the list from each school who were present in the school on the day enumerators conducted baseline exams were selected into the sample.[20] These 1900 students from the 38 study schools comprise our baseline sample. Just over half of the students are female and the mean age at the beginning of P1 is 7 (Appendix Table A3, columns 1-3).

3.2 Learning Outcomes

We assess student learning using a set of exams conducted at the beginning and end of the school year that tested first-grade students on their ability to read and write Leblango. 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 late November 2013. Exams were administered 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.

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: recognizing letters of the alphabet, reading simple words, and understanding sentences and paragraphs" (RTI International, 2009); the exam has been adapted to dozens of languages and implemented in nearly 70 countries around the world. We use a version of the EGRA adapted to Leblango that was developed and used in Uganda in 2009 as part of an assessment of the reading ability of 2000 students in 50 schools across the country (Piper, 2010). The exam covers six components of reading ability: letter name knowledge, initial sound

[20] These students are likely in the upper part of the distribution of family socio-economic status, given they were selected conditional on attendance and enrollment early in the school year; this could affect the generalizability of our results. There is no evidence of differential enrollment based on treatment group and parents were not informed of Mango Tree's involvement with schools until well into the first term.

identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The first four components involve students attempting to read letters, sounds, and both real and invented words. The last two components have students attempt to 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, making use of visual aids. The examiners marked each question correct or incorrect during the exam.

Writing Leblango

To capture students' ability to write, we use a writing assessment designed by Mango Tree and previously used to monitor writing skill acquisition in their pilot testing. Writing tests were conducted in a group setting with a single examiner handing out materials and instructing pupils to write a story. These assessments were scored off-site by a native speaker and expert in Leblango writing acquisition. In the first section of the test, students are 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. Given names come from a small list of names with known English spellings. Each name was scored separately in two categories: spelling and capitalization. In the second section of the test, students were asked to write or draw a story 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.[21] Each writing concept is scored on a 5-point rubric: 1. Experimenting, 2. Emerging, 3. Developing, 4. Capable, and 5. Experienced.

Combined Exam Score Indices

Our main learning outcomes – reading and writing in Leblango – are measured using the endline exams. Each exam contains 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

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

questions; some are scored based on a student's speed while others are untimed. We present the effects on each module separately, as well as construct combined outcome indices using principal components analysis (PCA) to measure overall reading and writing ability.[22] 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 the endline control-group standard deviation. This gives the index 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. 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 provides more information about the control group's progress over the year.

Balance and Attrition

Of the students tested at the baseline, 78% were also tested at the endline. This gives us a longitudinal sample of 1481 students, which is 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 among those who were lost to follow-up (attritors). The baseline sample is balanced in terms of basic demographics and the PCA indices of overall reading and writing ability (columns 1-3). Although some of the individual exam components show correlations with study arm at baseline, these differences even out when the exams are combined into the overall indices. Our main regression specifications control for baseline scores, which should address any imbalance on observables. In general, student characteristics do not correlate with attrition across study arms (Appendix Table A3, columns 4-9).

[22] 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.

3.3 Classroom Observations

In addition to the baseline and endline examinations at each school, enumerators collected classroom observations three times during the school year. These visits took place in July (term 2), August (term 3), and October (term 3) and involved observing two 30-minute lessons in each of the school's two P1 classrooms. The observations captured information about teaching strategies, student behavior and engagement, discipline, language of instruction, and the focus of each lesson.

Classroom observations were collected in three 10-minute blocks of time. For each block, the enumerator recorded the start and end time and indicated whether a teacher engaged in a range of actions during the block of time. The enumerators also 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, such as doing the activity in a group or on their own, using a specific material such as a slate for writing or a reader for reading, and whether English or Leblango was used. The full classroom observation instrument is shown in Appendix Figure A1.

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 allows us to exploit the patterns of correlations between different variables in the classroom observations, and to study how these patterns vary across the two program variants. We conduct separate factor analyses for classroom management, reading pedagogy, and writing pedagogy, and retain 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 intentionally pool all three study arms when doing the factor analyses, because we think that the changes in teacher and student behavior caused by the treatment will yield different factors and factor loadings. The resulting factors and factor loadings are shown in Appendix Tables A4, A5, and A6; we give descriptive names to each factor based on the specific behaviors that load on that factor.

We use three factors to 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. These explain 0.81, 0.31, and 0.25 of the variation in observations, respectively.

Our procedure retains five factors for 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; "Basic Elements in Breakout Sessons," which involves students working in smaller groups; "Leblango Sentences in Reader"; and "Paragraphs in Primer". These factors explain 0.49, 0.35, 0.27, 0.19, and 0.15 of the variation, respectively.

Lastly, the factors for writing pedagogy are "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". These explain 0.46, 0.31, 0.21, 0.16, and 0.12 of the variation across writing activities respectively.

3.4 Empirical Strategy

Effects of the Programs 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. While the treatment was assigned at the school level, our main analyses focus on student-level outcomes. 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 in the full-cost and modified version of the program respectively,with the omitted category being in 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 modified programs, respectively, on exam scores. 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. 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).[23] Our preferred specifications also control for the baseline value of the outcome variable,$y_{is}^{baseline}$, as specified in our pre-analysis plan and to address any potential baseline imbalance on test scores.[24] We also show our results without baseline controls. To account for the fact that the treatment was randomized at the school level, we uniformly report robust standard errors that are clustered by school.

Identifying Mechanisms and Mediators

In addition to measuring the effect of the two programs on test scores, we examine effects on student and teacher behavior using the classroom observations data. We analyze the 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.

To examine the extent to which changes in these classroom behaviors mediate the effect of the NULP on student test scores, we run the Acharya, Blackwell, and Sen (2016) sequential *g*-

[23] The probability of assignment to each study arm varies by stratification cell because two cells contained just two schools instead of three.
[24] https://www.socialscienceregistry.org/docs/analysisplan/36/document

estimator. This estimator improves upon the common strategy of studying mechanisms by controlling for intermediate outcomes in a regression analysis, which can lead to arbitrarily biased estimates and results that have no straightforward interpretation.

We focus on the comparison between the full-cost and reduced-cost versions of the NULP (their estimator cannot handle multi-valued treatments). We merge the classroom observation data into exam score data by teacher, and restrict the data to the full-cost and reduced-cost study arms. Our preferred specification uses the factor analysis indices describing classroom management, reading, and writing pedagogy as mediators although the results are substantively identical if we use raw classroom observation variables.

Mediation analysis involves estimating what proportion of the treatment effect is explained by mediators – variables affected by the treatment that in turn influence the main outcome. The mediators we examine here are the classroom management and pedagogy factors estimated from our classroom observation data. We first re-center all the mediator variables (i.e., the factor indices) relative to the reduced-cost program, by subtracting off the reduced-cost program mean. We then estimate the following regression:

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

The notation follows equation 1, but also includes $\boldsymbol{M}'_{iso}$, a vector of mediator variables, where *o* indexes a specific observation block in classroom *c* and school *s*. We allow the effect of the mediators to vary across study arms by including an interaction term. To consistently estimate τ and λ, we need to satisfy a "no intermediate variable bias" assumption – that there should be no variables omitted from our regression that are affected by the treatment and influence the outcome that are 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, we include *all* the mediators from the classroom observations in this regression. Second, we also control for a vector of other intermediate variables $\boldsymbol{I}'_{iso}$ that could be intermediate confounders: fixed effects for the block of the classroom observation, the round of the visit, the day of the week, and the enumerator who did the visit, and a control for the total number of observation blocks for a given

classroom observation.[25] 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 the product of the mediators and the estimated coefficient from the raw outcome (y_{isc}) and likewise for the interaction term:

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

The result, $y_{isc}^{demediated}$, is the outcome variable after purging the effects of changes in the mediator variables. We can use it to run a modified version of equation 1, regressing the de-mediated value of *y* on the treatment indicator and our baseline controls:

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

Acharya, Blackwell, and Sen (2016) show that under the right conditions, equation 5 estimates the *average controlled direct effect* – the effect of the full cost treatment on test scores, relative to 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.

Assessing the Predictive Power of the Classroom Observations for Student Learning

Standardized classroom observation measurements are strong predictors of student

[25] The overwhelming majority of classroom observations were 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.

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 and other similar standardized tools in that it focuses more on objective behaviors than on subjective assessments of teaching quality.[26] Since our tool was designed for this study and has not been validated in other contexts, we conduct two supplementary analyses to assess its predictive power for endline test scores.

First, we collapse the data to classroom-level means and run simple regressions of test scores on all our classroom observation variables, estimating:

$$\bar{y}_{sc} = \beta_0 + \bar{\boldsymbol{M}}'_{sc}\varphi + \epsilon_{sc} \quad (6)$$

where $\bar{y}_{sc}$ is the average endline exam score in classroom *c* in school *s* and $\bar{\boldsymbol{M}}'_{sc}$ is a vector of the average values of the mediators in classroom *c* in school *s*. We use the R-squared from this regression to assess the predictive power of using the mediator variables separately and linearly to predict endline exam scores.

Using the mediators as separate linear predictors is a restrictive assumption: real education production functions are likely to involve areas of both increasing and diminishing returns as well as complementarities between inputs. We would quickly encounter dimensionality problems if we simply added higher-order terms and interactions to our regression, since there are only 78 classrooms in our dataset. Instead, we use a machine-learning approach to assess the predictive power of our mediator variables. We apply kernel regularized least squares approach of Hainmueller and Hazlett (2014) to the same basic model represented in equation (6), allowing the estimator to select the interactions and higher-order terms. This estimator converges to the true R-squared asymptotically.

In small samples, Hainmuller and Hazlett show that this estimator may be biased upward. To get a sense of the empirical importance of this potential bias, we apply the estimator to random noise. We replace the real mediators with random numbers, using the same number of random

[26] The CLASS does have assessors score actions at a fairly granular level. However, Araujo et al. (2016) note that these item scores tend to be highly correlated and reflect underlying teacher quality; as a result they use the overall average score on the CLASS rather than examining the individual components.

variables as we have mediators in the real data (26 total random variables). 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.

4 Results

4.1 Learning During P1 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 are P1 unable to recognize a single letter of the alphabet, and the majority of those students leave P1 having made no progress whatsoever (Figure 1, 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.[27]

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.[28]

The measured increases in exam scores in the control group form a natural basis for comparison for the effects of the two versions of the NULP on exam scores. In the absence of the intervention, students improve by 0.15 SDs in reading over the course of first grade, and 0.47 SDs

[27] The maximum raw score on the letter name-knowledge section of the EGRA is 100 letter names correct (some letters are repeated). However, consistent with the EGRA protocol students who did not get any answers right in the first ten letter names were skipped ahead to the next section to minimize embarrassment and discomfort. Thus a zero score on this section of the exam indicates that the student got no answers correct out of the first ten.

[28] 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.

in writing. Our randomized evaluation measures the additional gains caused by the program; we can compare those additional gains to the typical gains experienced by a child during P1.

4.2 Effects on Reading

The program had a substantial effect on reading scores, illustrated in Panel B of Figure 1. It shows the endline distribution of letter name knowledge scores by study arm. The full-cost version of the NULP reduces the proportion of students who cannot recognize a single letter by nearly half, and more than triples the share that can recognize 21 or more letters per minute. The effects are similar if we instead examine the total number of points scored on the EGRA (Figure 2 Panel B). The reduced-cost version of the NULP achieves smaller improvements in both letter name recognition and overall EGRA performance. It shifts the score distribution to the right, but does so by a smaller degree than the full-cost variant.

The estimates of the impacts of the two versions of the NULP on EGRA scores are shown in Table 2, which estimates equation 1.[29] Column 2 presents the impact on students' knowledge of letter names, the principal learning goal that Mango Tree sets for P1 students. The full-cost version of the program has a very large impact on letter name knowledge: scores increase by 1.01 standard deviations. The modified program improves performances in recognizing the names of letters by 0.41 SDs, which is still a significant gain but less than half as much as the full-cost version of the program. The difference between the effects in the full-cost and reduced-cost programs is 0.51 standard deviations and is statistically significant.

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 that are uniformly significant at the 0.05 level. The modified program, however, has no statistically-significant effect on any EGRA component other than letter name knowledge. The reduced-cost version of the program, then, improved only the headline measure of literacy emphasized by Mango Tree, with no benefits to other, more advanced aspects of literacy.

This finding is verified by Column 1 of Table 2, which presents estimates for the combined reading score index. The full-cost program raises this index by 0.64 SDs, confirming that the large

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

effect of the program on exam scores is not merely an artifact of focusing on knowledge of letter names. Even taking 0.64 SDs as our best estimate of the program's impact on reading ability (rather than the effect of the program on letter name knowledge), the effect of this program is among the largest ever measured in a randomized trial of an education program (McEwan, 2014). Moreover, we can reject gains smaller than 0.37 SDs at the 0.05 level; in the few cases where randomized evaluations of education programs have found large effects, those estimates have been paired with wide confidence intervals that do not exclude much smaller impacts. In contrast to the large gains in the full program, the modified program's effect on the EGRA index is just 0.13 SDs and is statistically indistinguishable from zero.

4.3 Effects on Writing

We present the effect of the two versions of the program on writing ability in Table 3. [30] Columns 2 and 3 show that both versions of the program have large effects on the first section of the exam, which asks students two write their first and last names. The full-cost program also has positive effects on the second section, in which students are asked to write/draw a short story (Columns 4 to 10). The combined writing test index rises by 0.45 SDs (Column 1), which is statistically significant at the 0.05 level.

The modified program, however, has uniformly negative effects on the story-writing component of the exam, with the negative effects on Voice, Word Choice, and Presentation reaching significance at the p=0.05 level. The combined writing test score index falls by 0.16 SDs, although this drop is not statistically significant. Taken together, the results suggest that the modified version of the program significantly boosted the headline measure of name writing, at the cost of progress in overall writing skills, and in particular the ability to actually write a passage.[31]

4.4 Cost-effectiveness

[30] Just as for reading, the writing test results are essentially unchanged by omitting the baseline exam score controls (Appendix Table A8).

[31] 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).

The large effects of the program naturally raise the question of its cost-effectiveness. While few other programs have shown such large gains, can the NULP compete on a value-per-dollar-spent basis? We examine this question in Table 4, which presents the total cost per student of each program version, as well as the cost per 0.20-SD gain and the SD gain per dollar spent for three different measures of the program's effectiveness. The full-cost program cost $15.39 per student while the reduced-cost version cost $6.05 per student.[32] The full-cost version is at the 93rd percentile of the cost distribution for the studies covered by McEwan (2015), while the reduced-cost version is at the 63rd percentile.

We first present the cost-effectiveness results using the estimated program effects on letter name knowledge, the most important outcome emphasized by Mango Tree for P1 students. Using this measure, the two versions are relatively comparable, both increasing letter name knowledge by 0.07 SDs for each dollar spent. The full-cost program is a bit more costly per student per learning gain – it would cost an extra 6 cents per student to raise letter name knowledge by 0.2 SDs. Using letter name recognition to measure cost-effectiveness, the reduced-cost program is roughly as cost-effective at raising letter name knowledge – and 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 instead of just letter-name knowledge reverses this conclusion: the full-cost version of the program yielded over twice the gains in performance per dollar compared to the reduced-cost version (0.04 standard deviations per dollar vs. 0.02 standard deviations 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 cost version.

Using effectiveness estimates from the writing ability index shows an even starker pattern: because the reduced-cost version of the program actually 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 will *decrease* writing performance by 0.03 SDs.

These results raise general questions about the use of cost-effectiveness measures in comparing the effects of education programs: they may mask considerable heterogeneity in program impacts across educational domains or measures, leading to apparently-cheap gains that come at potentially-large hidden costs. These findings suggest that policy makers should be

[32] These figures are based on actual expenditures in 2013 and include only incremental program costs, excluding costs related to materials development, curriculum design, etc.

cautious about using cost-effectiveness results to make policy decisions. Summaries of cost-effectiveness typically focus on a single test score measure, often because the underlying studies look at just one outcome variable (JPAL 2014, McEwan 2015). This approach may mask important heterogeneity in cost-effectiveness across different outcome measures.

These calculations also suggest a potential trade-off between a program's cost-effectiveness and affordability. Two programs may be equally cost-effective producing learning gains at similar costs, yet one may be prohibitively expensive to provide to a large number of schools, teachers, or schools. It is essential to consider the absolute cost of a program, in addition to cost per learning gain.

5 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 that are targeted as goals for P1 students – letter recognition and name writing – with no gains in other areas and statistically-significant losses on more advanced aspects of writing ability. The two variants of the program were randomly allocated as complete packages, so we cannot causally separate the benefits of each individual part of the program. Instead, we exploit detailed classroom observation data to examine how the program affected teacher and student behavior in the classroom, and test the role of those changes in mediating the effects of each version of the program. Using these results and knowledge about how the program operates, we discuss potential reasons why the results of the NULP are so sensitive to minor changes in implementation. In particular, there is evidence that certain program components function as strong complements to one another, so removing one can lead to sharply different results.

5.1 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, speaking/listening, and the proportion of time where the local language is used. Teachers in both

treatment groups 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, and this difference is marginally statistically significant ($p<0.10$). Teachers in the full-cost version of the program actually spend less time (3.6 percentage points less, $p<0.05$) on writing than the control group. Given relative the impacts on writing, this suggests that time spent learning to write was much more productive in the full-cost NULP compared with the reduced-cost version or the control group. We estimate that students in the full-cost program gained 0.011 SD in writing scores for every hour spent learning writing, as opposed to 0.002 for the control group and 0.004 for the reduced-cost group.

Both versions of the program use Leblango more often in the classroom than the control group does. The difference between the use of English in the full-cost and reduced-cost versions of the program is just 3.2 percentage points (and not statistically significant), and the control group already uses Leblango 69% of the time. This sheds some light on the relative importance of mother-tongue instruction in generating the results of the NULP. Given the high base rate of mother tongue instruction, and the fact that the program effects are very different between the two program, it seems unlikely that the mother-tongue focus of the NULP is a key determinant of its effectiveness.[33] However, it may be an important complement to other inputs, which is a possibility we will return to below.

5.2 Classroom Management and Pedagogy

We examine how time is allocated within reading and writing classes in Tables 6 through 9 using raw classroom observation variables and combined factor indices summarizing all of the variables in the dataset and how they co-vary with one another. Tables 6 and 7 present the results for reading activities. Students in the full-cost NULP are more likely to spend time reading from readers and primers – materials that the NULP provides to classrooms (Table 6, columns 5 and 6). Although full-cost and reduced-cost classrooms both received the same primers and readers, we see a more-muted effect on material use for the reduced-cost program, and the treatment effect is

[33] Given the drop in the use of English 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 0.3 SD for the full-cost study arm (Appendix Table A10).

statistically insignificant for the use of readers. The control group hardly uses these types of materials at all – just 3% of the time for primers and 6% for readers – reflecting the extremely low availability of classroom materials in the region.

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 significant, with more focus on sounds in the full program. The three study arms have no detectible differences in practicing letters and words. However, based on exam score measures, students in the full-cost program perform much better on these aspects of writing. This suggests that control-group students get little benefit from their class time spent on letters and words. One possibility is that the control group moves too quickly through the curriculum, and does not spend enough time building a foundation of basic skills upon which more-complicated skills can be built (Pritchett and Beatty 2015). The pedagogical design of the NULP directly targets this issue by having teachers move more slowly through the basic reading skills, covering only half of the Leblango phonemes in grade 1.

The impacts on the factor analysis indices in table 7 reveal more subtle patterns. Classroom management during reading classes differs somewhat across the study arms: both full and reduced-cost program teachers spend somewhat less time on keeping students focused, and this is significant at the 0.10 level for the reduced-cost program. This pattern may be the result of students being more engaged, and thus requiring less intervention to keep them on-task. We also see different pedagogical strategies across study arms. There is an increase in practicing reading Leblango sentences from readers, and this change is significantly larger for the full-cost classrooms (Column 7). Both full- and reduced-cost program classrooms spend more time reading paragraphs out of the primer (Column 8). These activities replace two strategies used in control group: whole-language exercises at the chalkboard, where the class goes through all concepts from letters to paragraphs together; and working on basic elements in small groups. These declines only reach statistical significance for the full-cost classrooms (Columns 5 and 6).

Tables 8 and 9 present the result from the classroom observations during writing activities. Both full and reduced-cost groups spend more time on name-writing (Table 8, column 5). The full-cost treatment group spends less time than the control group on writing letters, with all other elements unaffected. Crucially, the reduced-cost group spends *less* time than the control group on writing sentences. This is consistent with their declines in performance on the story-writing

component of the writing test (Table 3). Some of the differences in writing scores across study arms can be explained by the use of materials. Full-cost program classrooms are much more likely to practice writing on slates, which substitute for paper. Reduced-cost program classrooms do no exhibit these changes, and instead spend more time on "air-writing" – tracing out the shapes of letters in the air. Slates are essentially unused in both reduced-cost program and control classrooms, consistent with their general scarcity in the region's schools (recall that slates were only provided in the full-cost program classrooms). The full-cost program classrooms spend much less time copying the teacher's text, and much more writing their own text.

The pedagogy factor analysis indices for writing in Table 9 tell a similar story. The index for practicing Leblango using slates is massively higher in the full-cost program classrooms (Column 6). It increases somewhat for reduced-cost classrooms as well – reflecting the fact that this index loads on multiple underlying variables, and can be positive even in the absence of slates. The smaller estimates here likely reflect the reduced-cost teachers attempting to carry out the pedagogical strategy pushed by the NULP, but achieving limited success because they lack a key input. Both treatment arms show drops in copying teacher's text; this effect is again significantly larger for the full-cost group (Column 5). The full-cost group also has lower values for the index for drawing pictures and letters on paper, with high levels of energy and participation; class activities that combine pictures, words, and stories become more common instead (Columns 4 and 7). Classroom management during writing classes differs significantly across study arms. Teachers in both treatment arms spend less time keeping students focused, suggesting that children may be more interested in what is being taught (Column 1). Both treatment arms also exhibit increases in having a solid lesson plan (statistically significant for the reduced-cost program) and the teacher being active throughout the classroom (statistically significant for the full-cost program) (Columns 2 and 3). The latter 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.

5.3 Mediation Analysis

How much of the differences in our treatment effects can be attributed to the differences in time allocation, classroom management, and pedagogy? To answer this question, we conduct a mediation analysis using the sequential *g*-estimator of Acharya, Blackwell, and Sen (2016). The

version of the sequential *g*-estimator described in Section 3.4, answers the question: what would be the effect of the treatment on the outcome if all the variables in the classroom observations were held at the values from the reduced-cost study arm? This allows us to summarize how much of the treatment effect can be explained by changes in the mediators. We run this estimator on a reduced sample that includes just the two treatment arms, because it does not allow for multi-valued treatments, and also because the key question we want to answer is why the full-cost program was so much more effective that the reduced-cost version.

Table 10 presents the results. We find that the changes in classroom observation mediators can explain only a small fraction of the difference in the treatment effects across study arms: 5.5% for reading (2.4% for letter name recognition alone) and 5.7% for writing with an adjusted R-squared of 0.191 and 0.304, respectively.

These results of our mediation analysis have three possible explanations. First, are these findings simply the result of poor measurement of classroom behaviors? We assess this possibility in two ways. First, we estimate a linear regression of classroom-mean test scores on classroom-mean mediators (equation 6). This yields an R-squared of 0.13 for reading and 0.28 for writing. Second, we estimate a kernel-regularized least-squares regression that allows for higher order terms and interactions without dimensionality problems (Hainmuller and Hazlett 2014). This estimator yields an R-squared of 0.82 for reading and 0.99 for writing. These figures are high, but we can verify that they are not the mechanical result of including a large number of predictors. Replacing the actual data with random noise yields R-squared values of 0.028 for reading and 0.034 for writing suggesting that any upward bias in our estimates of the predictive power of the mediators is minimal.

Second, the relevant changes in teaching may be too subtle to detect using our classroom observations instrument. Our tool can measure time use and specific actions taken by teachers and students, but cannot tell us how effective the time used and the actions taken are in terms of causing learning.

Third, the relevant changes in teaching may be combinations of a variety of different complementary behaviors. The machine learning approach was able to achieve a much greater predictive power for test scores than ordinary least squares on the same data, because it considers nonlinear terms and interactions between predictors. The correct functional form for the mediators of the treatment effects may include these higher-order terms and interactions.

5.4 Potential Complementarities between Inputs

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 they may be due to strong complementarities between inputs in the learning process for reading and writing. These complementarities mean that removing specific inputs can drastically change the returns to investment in certain domains of learning (see for example, Mbiti et al. 2017).

This story is easiest to understand for writing, because the reduced-cost version of the NULP did not include slates from the set of physical inputs provided 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 more likely to be practicing writing on their own. In contrast, reduced-cost version students (without slates) were more likely to practice writing using their hands "in the air," presumably one factor in reducing writing ability. While the slates were not randomly allocated to classrooms to test for complementarities, our results suggest strong complementarities between physical inputs (slates) and worker human capital (teacher training and support) for producing writing skills.

This story is more subtle than it appears, however. Slates are most useful for the simplest writing tasks: practicing letters, writing names, and practicing a single word. They are not ideal for practicing 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 the name-writing task, still improved substantially. Our detailed classroom observation data help to explain how this might have happened. Because the slates are an important complement to the other inputs in the NULP, basic writing skills were difficult to teach without them. Students in the reduced-cost schools thus took longer than the control group to master the basics of writing letters and words. As a result, they were unable to spend enough time practicing writing actual sentences.

While the reduced-cost NULP's negative effects on writing are striking, the magnitude of the decline in impacts – compared to the effects of the full-cost NULP program – is nearly as large for reading. Thus, complementarities between slates and the other NULP inputs cannot be the entire story for our results. Instead, details – some of which would not seem qualitatively important ex-ante – must be driving the differences in effectiveness across study arms. In our case, one key

detail is the amount of follow-up that teachers receive, which helps them stay "on course" with the program. The full-cost study arm receives over twice as many of these visits, where teachers get feedback on how they are doing and advice for how to improve, which is a common feature of successful education systems (Bruns and Luque 2015). Support visits can also serve as a form of monitoring to help ensure that teachers are actually implementing the NULP model. This is monitoring without associated incentives: just checking in with teachers to keep them on task, which has been effective in other education programs in Africa (Aker and Ksoll 2015).

There is evidence of this, from the classroom observations data, in which enumerators recorded the date of each visit and which lesson was being taught. In theory, primary schools in Uganda follow a national thematic curriculum with a specified theme and subtheme for each week.[34] We coded each classroom visit to indicate whether the class was on track or not – i.e. whether the correct theme and subtheme were being taught for the week. Just 29% of control-group classrooms were on track by this measure, as compared with 51% for the reduced-cost program and 61% for the full-cost program. The difference in the rates of being on track across the two NULP program versions is statistically significant at the 10% level for a basic specification, but the p-value rises to 0.3 if we include day-of-week fixed effects.

Although teachers in both the full-cost and reduced-cost NULP schools each received a teacher's guide to follow, which may have helped lessons to stay on track, this does not explain the difference between the two variants. Our analysis here captures compliance only at the extensive margin – whether the teacher is covering the correct lesson for a given day. Effects at the intensive-margin – whether the teacher is covering the material *correctly* – may be even larger. One possibility is that the additional support visits lead to a higher degree of compliance with the NULP curriculum and pedagogical framework. We hypothesize that this complementary input helps to reinforce the teaching practices and instructional model that teachers learn during their training lessons.[35]

Separate evidence that the NULP's benefits depend on a complex set of complementary inputs such as teacher behavior within the classroom comes from our mediation analysis. Despite

[34] The NULP curriculum was explicitly designed to follow the government theme and subtheme.

[35] Banerjee, Banerji, Berry, et al. (2016) find suggestive evidence that teachers tend to revert to their ingrained habits after receiving training: in the scale-up of a program in India teachers did not implement a new teaching methodology even though they attended (and approved of) training in the methodology. Additional training and support throughout the school year led to much higher rates of implementation, and greatly improved learning.

the significant differences between full-cost and reduced-cost schools in the classroom observations data, we can explain just 6% of the difference in treatment effects using those mediators linearly. The mediators do have decent predictive power in simple OLS regressions, but their predictive power is very strong using the kernel regularized least squares machine-learning estimator of Hainmueller and Hazlett (2014). 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 studying the role of the mediators also has these complementarities. Even if we were able to impose the correct functional form on the mediation analysis, however, it is not certain that this would shed light on the mechanisms behind the NULP's impact. The true production function is likely to include a large number of interactions and high-order terms that are not straightforward to interpret.

6 Conclusion

In this paper we compare two delivery models for a primary literacy program, randomly assigned to schools in northern Uganda: a full-cost model delivered by the organization that designed the program, and a reduced-cost model delivered through a training-of-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. Overall reading improves by 0.64 SDs and overall writing by 0.45 SDs. These effects on learning are among the largest ever measured for an education intervention. Focusing on the simpler literacy skills, we see gains around 1 SD for letter recognition and writing one's name. The reduced-cost version of the program fares much worse. It improves only the simplest reading and writing outcomes, leaving advanced reading skills unchanged and actually worsening students' advanced writing skills relative to control.

These qualitatively different outcomes arise from seemingly-minor differences in implementation details. Students in the reduced-cost version of the program were taught by the same types of teachers, using the same curriculum, and with the same materials as those in the full-cost version – but experienced reading gains that were 80% smaller, and writing gains that were 135% smaller (that is, negative). Objective comparisons of the two program variants confirm that the differences are small: the two program variants differ by only 6% on the Arancibia, Popova, and Evans (2016) measures of the attributes of in-service teacher training programs. The reduced-cost program is also quite policy-relevant. It was designed specifically to emulate how the program

might be scaled up, by modifying some inputs and eliminating others. These changes make it quite attractive from a cost perspective, carrying a 60% lower price tag than the full-cost NULP.

Our study contributes to a growing literature documenting that the effectiveness of educational interventions can be extremely sensitive to small changes in implementation. Duflo (2017) argues that in a wide range of contexts, the "plumbing" of programs – details that do not seem economically important – can matter a huge amount for the practical results of a program. Nadel and Pritchett (2016) argue that when the design space is extremely "rugged" (so small changes in design details cause large changes in outcomes), randomized trials lack construct validity. This is a deeper issue than external validity: even if a program would work equally well outside of the study setting, we may not be studying the same underlying object that will 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.[36] Our results verify and extend Bold et al.'s 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 classroom and so it unclear what exactly drives the difference in effectiveness across delivery modes. We use detailed classroom observations to examine how teaching differs across study arms and also to conduct mediation analyses to measure the importance of these differences for explaining the final outcome.

How exactly did the fairly subtle (if financially appealing) changes in the reduced-cost version of the lead to such large differences in results? Using both detailed knowledge of the

[36] 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.

program and the results from our evaluation, we argue that the sensitivity of the NULP's results to program details can be explained by powerful complementarities between inputs and the fact that learning is path-dependent. One potential explanation for the decline in writing results in the reduced-cost program points to the elimination of slates. This led to slower-than-usual progress in students developing basic writing skills, as teachers attempted to implement the curriculum without a key input. Students thus spent less time on actually writing sentences, and their performance on the advanced writing tasks on the endline exams suffered. It is also likely that the quantity and quality of teaching support visits complements other aspects of the program; reducing the number of visits thus sharply reduced the benefits of the rest of the program. Our machine learning results support the idea that complementarities between inputs are crucial. Just 6% of the performance gap between program variants can be explained using our classroom observation mediators. OLS regressions of outcomes on mediators show higher predictive power, but we can still explain less than a third of the classroom-level variation in the data. Allowing for nonlinearities and interactions, however, allows us to explain an extremely large fraction of the variance in test scores.

Our results contribute to an ongoing debate about the validity of drawing inferences from experiments in economics. An extensive literature has criticized randomized experiments as being limited in their ability to guide policy and provide generalizable insights.[37] We provide evidence that construct validity is another important limitation of randomized experiments. We also shed light on why construct validity is so limited in educational interventions: complementarities and path-dependence in the education production function. 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 on complementarities is limited. Mbiti, et al. (2017) find evidence in favor of complementarities while List, Livingston, and Neckermann (2013) do not. Few studies can even detect complementarities, let alone isolate their importance: the McEwan (2015) meta-analysis of education experiments in developing countries finds that 75% of experiments have only a single treatment arm, making it impossible to study complementarities

[37] For instance, see Deaton (2010) and Alcott (2015) 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.

between interventions.

The importance of complementarities in education production may help explain the limited benefits of most education programs. Three meta-analyses of hundreds of studies find average effects on test scores of less than 0.2 SDs, 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 systematically underestimate the total gains from investing in schools.

Complementarities between inputs and the variety of outcomes available create a significant challenge for assessing the mechanisms behind a specific program's impacts. Our experiment provides an extreme example, where the impact of a program is highly sensitive to the exact input mix used and depends on the outcomes we use to evaluate it. Given this sensitivity, and the huge variety of options policymakers have in implementing a program, determining the exact mechanisms behind a program's success may be infeasible. The same problem means that generalizing from a successful pilot to the results of a scaled-up version of the program may be impossible: 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.

Our results suggest that, in many cases, the design space for education programs truly is rugged, and altering the implementation of a program can qualitatively change what it does. Programs must therefore be evaluated continuously – not just during the pilot testing phase, but throughout the scale-up and implementation process – in order to accurately measure their effectiveness. The findings in this paper also highlight the trade-off between a program's cost-effectiveness and affordability. Two programs may be equally cost-effective producing learning gains at similar costs, yet one may be prohibitively expensive to provide to a large number of schools, teachers, or schools. It is essential to consider the absolute cost of a program, in addition to cost per learning gain.

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

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

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.

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

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., & 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

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

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.

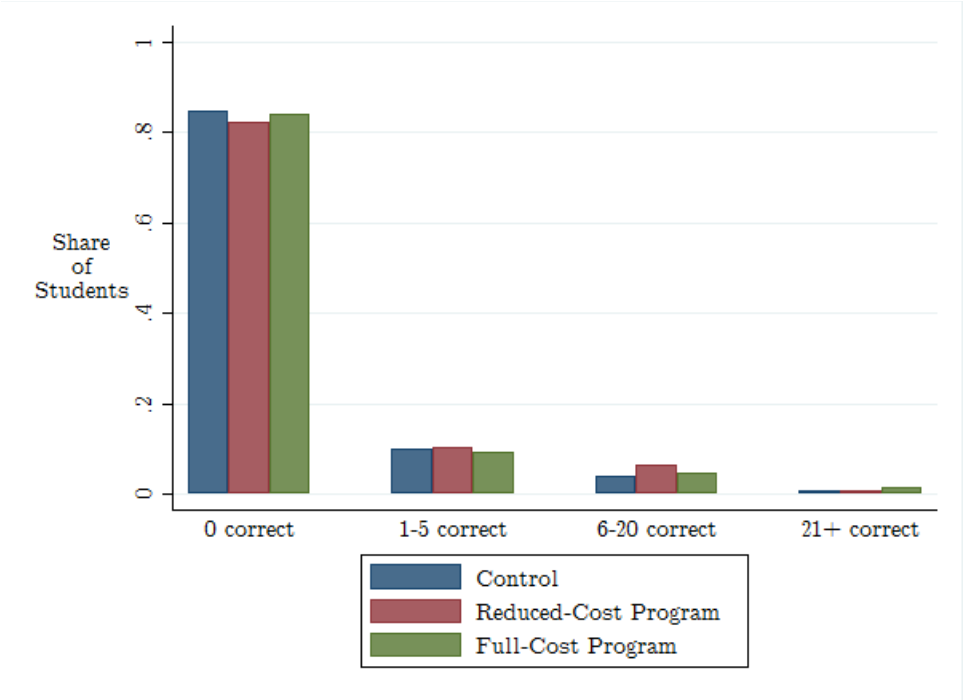
Tan, J.-P., Lane, J., & Lassibille, G. (1999). Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments. *The World Bank Economic Review*, *13*(3), 493–508.

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

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

Figure 1
Performance on Letter Name Recognition by Study Arm
(Number of Letters Correctly Recognized)

Panel A: Baseline



Panel B: Endline

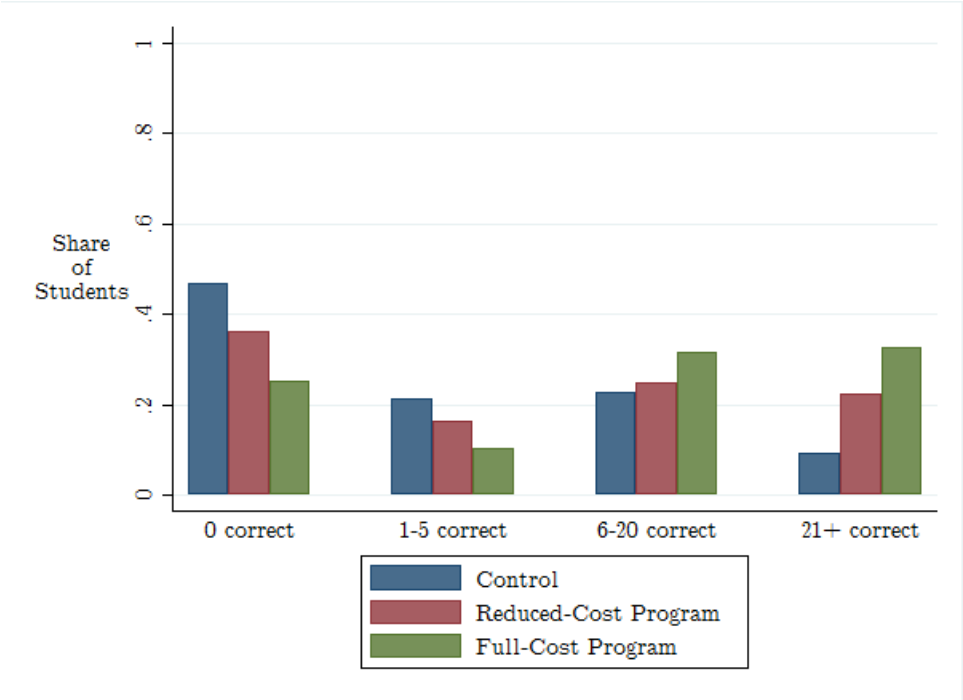
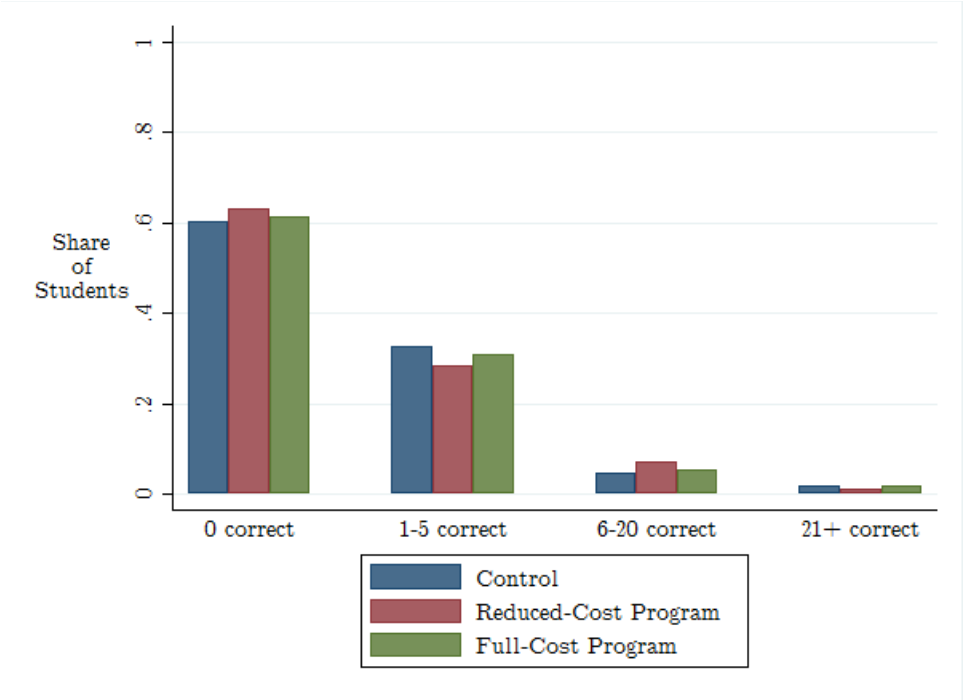


Figure 2
Performance on Overall EGRA by Study Arm
(Number of Letters Correctly Recognized)

Panel A: Baseline



Panel B: Endline

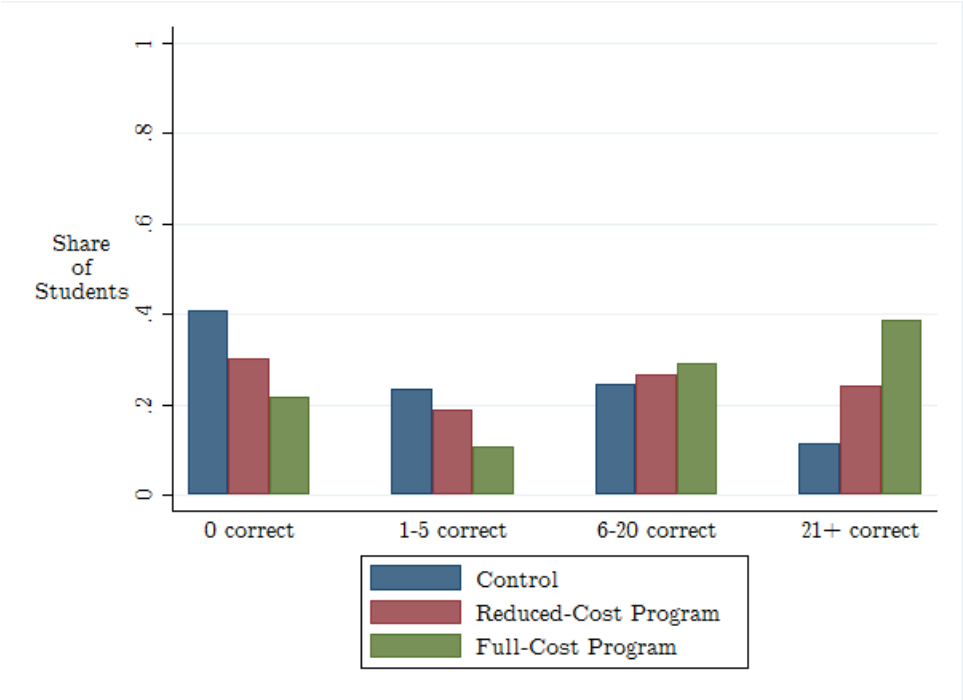


Table 1

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)	(2)	(3)	(4)	(5)	(6)	(7)
	PCA Leblango EGRA Score Index†	Letter Name Knowledge	Initial Sound Recognition	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Full-cost program	0.638***	1.014***	0.647***	0.374***	0.215**	0.476***	0.445***
	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.128)	(0.113)
Reduced-cost program	0.129	0.407**	0.076	-0.002	0.031	0.071	0.045
	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
Number of Students	1460	1476	1481	1474	1471	1467	1481
Adjusted R-Squared	0.149	0.219	0.057	0.066	0.075	0.074	0.058
Difference between full-cost and reduced-cost treatment effects	0.509***	0.607***	0.570***	0.376***	0.184*	0.405***	0.400***
	(0.127)	(0.159)	(0.128)	(0.092)	(0.093)	(0.117)	(0.12)
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	1.92	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; * 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†	African Name (Surname)	English Name (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
	(0.144)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
Reduced-cost program	-0.159	0.435***	0.450***	-0.274*	-0.316*	-0.313**	-0.262**	-0.330*	-0.253	-0.330**
	(0.122)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
Number of Students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Adjusted R-Squared	0.352	0.24	0.057	0.174	0.304	0.177	0.2	0.302	0.164	0.171
Difference between full-cost and reduced-cost treatment effects	0.608***	0.487***	0.861***	0.436***	0.757***	0.465***	0.437***	0.713***	0.474***	0.469***
	(0.128)	(0.135)	(0.154)	(0.148)	(0.173)	(0.118)	(0.139)	(0.174)	(0.151)	(0.115)
Raw (unadjusted) values§										
Control Group Mean	0.482	0.593	0.35	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control Group SD	1	0.685	0.533	0.372	0.594	0.393	0.416	0.59	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; * 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.

Table 4

Cost-Effectiveness Calculations

	Program Variant	
	Full-cost	Reduced-cost
Cost per student	$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***
	(0.014)	(0.017)	(0.013)	(0.036)
Reduced-cost program	0.054***	-0.003	-0.051***	0.079**
	(0.014)	(0.016)	(0.014)	(0.039)
Number of Observation Periods	1288	1288	0	1285
Adjusted R-Squared	0.078	0.063	0.131	0.126
Difference between full-cost and reduced-cost treatment effects	0.011	-0.033**	0.022*	0.032
	(0.015)	(0.016)	(0.012)	(0.029)
Control Group Mean	0.321	0.245	0.426	0.688
Control Group SD	0.278	0.32	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; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

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**
	(0.026)	(0.034)	(0.026)	(0.047)	(0.039)	(0.030)
Reduced-cost program	0.073***	0.053	-0.014	0.002	0.110***	0.040
	(0.025)	(0.035)	(0.029)	(0.054)	(0.037)	(0.030)
Number of Observation Periods	893	893	0.057	893	893	893
Adjusted R-Squared	0.054	0.025	0.045	0.044	0.091	0.222
Difference between full-cost and reduced-cost treatment effects	0.042*	-0.032	0.017	0.071*	0.057	0.022
	(0.023)	(0.033)	(0.025)	(0.036)	(0.040)	(0.026)
Control Group Mean	0.075	0.22	0.855	0.467	0.028	0.059
Control Group SD	0.264	0.415	0.353	0.500	0.165	0.237

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; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

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***
	(0.089)	(0.053)	(0.066)	(0.063)	(0.074)	(0.053)	(0.059)	(0.049)
Reduced-cost program	-0.160*	0.018	-0.044	0.123*	-0.076	-0.050	0.162**	0.121**
	(0.084)	(0.053)	(0.054)	(0.069)	(0.061)	(0.065)	(0.060)	(0.050)
Number of Observation Periods	890	890	0.057 890	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 full-cost and reduced-cost treatment effects	0.023	0.042	0.093**	-0.027	-0.239***	-0.052	0.100*	0.040
	(0.098)	(0.045)	(0.041)	(0.055)	(0.074)	(0.06)	(0.054)	(0.043)
Control Group Mean	0.185	0.051	-0.085	-0.094	0.187	0.067	-0.190	-0.069
Control Group SD	0.591	0.551	0.547	0.656	0.536	0.564	0.546	0.453

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; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 8

Classroom Observations: Writing Elements and Materials

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Element of Focus					Materials Used				
	Pictures	Letters	Words	Sentences	Name	Air Writing	Copying Text from Board	Writing Own Text	On Slate	On Paper
Full-cost program	0.097	-0.103*	0.031	-0.036	0.250***	-0.035	-0.202***	0.327***	0.319***	-0.194***
	(0.062)	(0.058)	(0.064)	(0.066)	(0.064)	(0.031)	(0.063)	(0.064)	(0.054)	(0.045)
Reduced-cost program	0.133**	0.045	0.149**	-0.142***	0.148***	0.058*	-0.097	0.070	0.027	-0.016
	(0.061)	(0.056)	(0.065)	(0.051)	(0.043)	(0.030)	(0.060)	(0.053)	(0.044)	(0.034)
Number of Observation Periods	539	539	0.057	539	539	539	539	539	539	539
Adjusted R-Squared	0.038	0.097	0.090	0.113	0.187	0.038	0.145	0.210	0.220	0.156
Difference between full-cost and reduced-cost treatment effects	-0.036	-0.149***	-0.118**	0.106*	0.102*	-0.093***	-0.105	0.257***	0.292***	-0.178***
	(0.044)	(0.053)	(0.048)	(0.056)	(0.057)	(0.028)	(0.062)	(0.053)	(0.055)	(0.051)
Control Group Mean	0.337	0.342	0.605	0.321	0.199	0.112	0.711	0.253	0.047	0.865
Control Group SD	0.474	0.476	0.491	0.468	0.401	0.317	0.455	0.436	0.212	0.343

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

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
	(0.104)	(0.058)	(0.065)	(0.083)	(0.091)	(0.096)	(0.058)	(0.065)
Reduced-cost program	-0.165	0.129**	0.055	-0.036	-0.178*	0.294***	0.034	-0.079
	(0.110)	(0.059)	(0.057)	(0.082)	(0.090)	(0.075)	(0.064)	(0.065)
Number of Observation Periods	537	537	$0.06	539	539	539	539	539
Adjusted R-Squared	0.086	0.25	0.188	0.107	0.213	0.294	0.087	0.227
Difference between full-cost and reduced-cost treatment effects	-0.010	-0.037	0.140***	0.232***	-0.295***	0.332***	-0.194**	0.060
	(0.106)	(0.078)	(0.046)	(0.077)	(0.084)	(0.078)	(0.073)	(0.056)
Control Group Mean	0.093	-0.016	0.157	-0.077	0.275	-0.339	0.062	0.026
Control Group SD	0.743	0.641	0.574	0.805	0.695	0.485	0.552	0.584

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; * 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 study arms	0.644***	0.665***	0.729***
	(0.219)	(0.182)	(0.180)
Adjusted R-Squared	0.262	0.191	0.304
Number of Observations	14524	14,343	13,587
Share of Treatment Effect Explained by Mediators	0.024	0.055	0.057
Raw (unadjusted) values[§]			
Reduced-Cost Program Mean	11.29	0.297	-0.043
Reduced-Cost Program SD	13.701	1	0.651

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; * p<0.05, ** p<0.01, *** p<0.001.