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

Jason T. Kerwin and Rebecca L. Thornton*

June 25, 2019

Click here for the latest version of this paper

Abstract: This paper demonstrates the acute sensitivity of education program effectiveness to the choices of inputs and outcome measures, using a randomized evaluation of a mother-tongue literacy program. The program raises reading scores by 0.64SDs and writing scores by 0.45SDs. A reduced-cost version instead yields statistically-insignificant reading gains and some large *negative* effects (-0.33SDs) on advanced writing. We combine a conceptual model of education production with detailed classroom observations to examine the mechanisms driving the results; we show they could be driven by the program initially lowering productivity before raising it, and also by missing complementary inputs in the reduced-cost version.

EconLit Subject Descriptors: I21, I25, O12, O15

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

1 Introduction

Children in sub-Saharan Africa are attending school more than ever before in history – but once in school, they learn very little (Boone et al. 2016, Pritchett 2013, Piper 2010). To address this learning crisis, hundreds of studies have rigorously evaluated the effectiveness of a wide range of educational interventions across a variety of contexts, countries, and types of programs.[1] Yet systematic reviews suggest enormous heterogeneity in effectiveness across studies, making it difficult to generalize from specific evaluations to inform policy (Nadel and Pritchett 2016). The variation in program effectiveness may be attributable to differences in context (e.g., India vs. Kenya) or the type of intervention evaluated (e.g., provision of materials vs. infrastructure upgrades), however the variation remains as large when comparing studies in similar contexts or the same type of intervention (Evans and Popova 2016, Vivalt 2017). The evidence of heterogeneity comes primarily from across-study comparisons, in part because most studies evaluate the effectiveness of a single intervention (McEwan 2015).[2] In contrast, this paper examines the variation of intervention effectiveness within a single study – holding both the context and the intervention type constant.

In this paper, we focus on two additional factors that affect the generalizability and policy relevance of education program evaluations: input choices and outcome measures. First, because every program differs in context, logistical constraints, and resources available, a common approach is to pick a highly-effective program and make it cheaper by modifying some of the most

[1] Evans and Popova (2016) discuss six systematic reviews of education program effectiveness in developing countries. Since their literature review, at least one additional review has been released (Glewwe and Muralidharan 2016).

[2] Notable exceptions include Bold et al. (2018) who test the effectiveness of NGO vs. government program delivery and Cilliers et al. (2019) who test different ways to deliver in-service teacher training.

expensive inputs. This option is appealing since effective interventions combine numerous inputs, many of which may seem unimportant. However, this strategy could lead to qualitative differences in program impacts if, for example, there are important complementarities between inputs. Second, there are many possible measures of learning: a wide range of tests, measuring a variety of skills and implemented in different languages. The variations in what is measured can play an important role in the interpretation of a program's measured effectiveness. In this paper, we demonstrate how these two issues can cause misleading conclusions about how to improve learning.

We use a randomized evaluation of a literacy program that was conducted in 38 government schools in the Lango sub-region of northern Uganda. The program, the Northern Uganda Literacy Project (NULP), is a mother-tongue-first early-primary literacy program developed by education and curriculum experts in Uganda. The NULP provides material inputs, high-quality teacher training, and support to first- to third-grade teachers. We compare primary schools that receive the program's entire array of high-quality inputs in their first-grade classrooms with schools assigned to a control group. The program is highly effective at improving mother-tongue literacy: after one year, it improves letter recognition by 1.01 SDs and improves overall reading by 0.64 SDs. The program also improves the ability to write one's first name by 1.31 SDs, write one's last name by 0.92 SDs, and overall writing performance by 0.45 SDs. These reading and writing effects are comparable to some of the largest measured in the literature.

Although highly effective, the program is costly for a developing-country education intervention, at about $20 per student per year. To study how reducing costly inputs would change the program's effectiveness, our experiment included a reduced-cost version of the NULP. This reduced-cost version made three changes to the program: 1) removing the most expensive material inputs; 2) a cascade model of delivery where teacher training and support was conducted by

government employees; and 3) fewer support visits to teachers. These changes amount to just a 6% difference on the Arancibia et al. (2016) indicators for in-service teacher-training programs, while reducing the per-student cost of the program by over 60 percent.

While the modifications to the program were relatively minor, these programmatic changes generate qualitatively different conclusions about its effectiveness. We find considerably smaller improvements in letter name knowledge in the reduced-cost version of the program than those in full-cost program schools (0.41 SDs). The reduced-cost version had no significant effects on more-sophisticated literacy skills (reading actual words or sentences), and gains to overall reading scores are small and statistically insignificant (0.13 SDs, p=0.327). The effectiveness of the two program versions diverge even further when we examine writing outcomes. The reduced-cost program shows gains only for the most basic tasks – the ability to write one's first name (by 0.45 SDs) and last name (by 0.44 SDs). At the same time, there are large, statistically-significant *negative* effects on the components that involved writing sentences (-0.33 SDs).[3] As measured by gains in letter name knowledge, the reduced-cost version of the program is slightly more cost-effective than the full-cost version (12% higher gains per dollar). For overall reading, however, the reduced-cost version is over 40% *less* cost-effective than the original NULP.

What led to the huge success of the original version of the NULP and why did the reduced-cost model fail? We present a conceptual model of an education production function, in which teachers maximize utility over multiple learning outcomes and the NULP affects learning by providing inputs and changing their productivity. The model can explain the backfiring effects of the reduced-cost program on advanced writing skills through several mechanisms. First, if the

[3] Other research has found unanticipated negative consequences of education interventions (Chao et al. 2015; Fryer and Holden 2012). Unlike those studies, however, the NULP provides no extrinsic incentives to students or teachers.

intervention raises productivity more in one skill than another, teachers may reduce investments in the second skill due to substitution effects. Second, a similar pattern can occur if there are important complementarities between inputs and one is omitted. Third, the program might reduce teachers' productivity in producing some learning outcomes, if, for example, teachers initially have to overhaul their teaching strategies and require practice with the new teaching methods in order to achieve later gains. It is possible the reduced-cost NULP never escaped this initial productivity dip for advanced writing skills.

We explore the implications of this model using a rich set of classroom observations, which suggest that the full-cost program resulted in large test score gains primarily through more-productive use of time and materials. For example, full-cost classrooms are 50% more likely than reduced-cost classrooms to use program materials during reading lessons (although this difference is not statistically significant). During writing lessons, there are large differences across study arms in the use of materials. Here, there are no significant differences between the reduced-cost program and the control, but large effects in the full-cost program. Students in the full-cost program shift from writing on paper to writing on slates, and rather than simply copying text from the board, they write their own text.

We find no evidence that the time devoted to reading and writing tasks is an important driver of our results. While teachers in the full- and reduced-cost programs spend 5-6 percent more time reading with students than those in the control, and spend 3-5 percent less time simply lecturing to students, there are no differences in these measures across the two study arms. We find a moderate increase in the use of local language (8-11 percent increase), but again, no difference across the two program variants.

While there were no large differences across the two program versions in amount of time

allocated to activities, we do find evidence of differences in of productivity. The returns to time on task (SD gains per hour of time spent) are 1.6 times higher than the control group in the reduced-cost program, and 4.5 times higher in the full-cost program. Similarly, the gains per hour of time spent on writing are 2.2 times higher in the full-cost program than in the control group. In contrast, the reduced-cost program teachers use writing class time *less* productively than the control group, achieving just 66% of the control-group gains on a per-hour basis. We also find changes in the way time is used: both versions of the program increase time spent on sounds and reading sentences, but the full-cost program increase is more than 50% larger for sounds (although this difference is not statistically significant) and over five times larger for reading sentences.

We also find that complementarities between inputs are also a likely mechanism for our results. We conduct mediation analyses – treating differences in classroom behavior as independent and linear predictors of learning – to explain the difference in effectiveness between the full- and reduced-cost programs. Using this method, we explain less than 4% of the difference in effectiveness across the two program variants, for both reading and writing. In contrast, machine-learning methods that allow for interactions and nonlinearities, predict far more of the variation in reading and writing scores than purely linear estimates: up to 18% of the difference in effectiveness in reading and 43% in writing. We show several different tests for overfitting.

Our results show that teachers in full-cost program schools were more engaged, and focused on different reading and writing elements during lessons. Teachers in reduced-cost program schools spent roughly the same amount time as full-cost teachers on reading and writing lessons, but ultimately used their time and material resources less productively. The negative effect of the reduced-cost program could be because those teachers – who did not receive more-intensive feedback and support visits – were unable to fine-tune their teaching to reach the higher levels of

productivity possible from the NULP model. Our results also suggest that complementarities may also play some role in driving the effects of the NULP on learning outcomes.

More generally, our findings argue for caution when modifying effective programs, even when those changes appear trivial. Indeed, we show that taking a highly effective program and cutting down on its costs may not just make it less effective, but may backfire, leaving some students worse off. Likewise, different learning metrics – often due to ad-hoc choices by researchers and partners – can drive vastly different conclusions about a program's effectiveness.

2 Context and Intervention

2.1 Context and the Northern Uganda Literacy Project

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

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

Because teaching effectively in African classrooms pose multiple challenges, the model

involves a carefully-designed bundle of inputs that directly address the challenges in rural Ugandan classrooms. We first describe the elements of the full-cost program. We then describe the reduced-cost version of the program and quantify the degree to which it differs from the full-cost version. The inputs provided to schools and their costs in each version of the program are listed in Table 1.

2.2 The Full-Cost Version

Uganda's official policy is that students in early primary school (grades one to three) are to be taught in their local language before transitioning to all English instruction in grade four. In practice, however, English is heavily used as the de facto language of instruction across the country. While it is important for students to learn English, full immersion in reading and writing a language that students do not yet know may also have powerful drawbacks (Webley 2006). Despite compelling theories for the benefits of mother-tongue instruction, well-identified evidence about its causal effects is sparse.[4] The NULP trains and supports teachers in literacy instruction in first grade, entirely in the students' mother tongue, Leblango. Teachers are instructed not to use written English on the board or in reading materials.

Primary school teachers in Uganda, who receive their basic training at teacher colleges across the country, receive additional training through the Teacher Development and Management System. The main approach follows a cascade model (i.e. "train-the-trainer"), in which trainers pass on skills and competences to government employees – Coordinating Centre Tutors (CCTs) – who then train teachers. In contrast, the NULP provides direct training and support to teachers using experienced Mango Tree staff (expert trainers), detailed facilitators' guides, and instructional

[4] Virtually all studies of mother-tongue instruction focus on Spanish-language immersion in the US (Rossell and Baker 2006). The only developing-country study we are aware of (Piper et al. 2016) finds that it improves mother-tongue reading scores by 0.3 to 0.6 SDs.

videos. Teachers undergo three intensive, residential teacher-training sessions on orthography and literacy methods, one before each of the three terms in an academic year. In addition to the residential trainings, there are six in-service training workshops on Saturdays throughout the year. CCTs undergo the same residential training sessions as NULP program teachers to become familiar with the NULP model; they also participate in in-service workshops.

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

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

Under both the status-quo government curriculum and the NULP model, students are exposed to fifteen half-hour literacy lessons per week. The government model divides these lessons into reading (5 lessons), writing (5 lessons), news (3 lessons), and oral literature (2

lessons). Under the NULP, lessons are divided into story-reading (5 lessons), creative-writing (5 lessons), and word building (5 lessons), with each lesson happening on each school day of the week.

Although schools receive capitation grants from the government to pay for instructional materials (e.g., books, chalk, wall charts, and teachers' guides), the material resources are often inadequate. To address this, NULP classrooms are provided a set of primers (textbooks that follow the curriculum and provide visual examples) and readers (books that provide text for reading practice). First-grade NULP classrooms are provided with slates that allow students to practice writing individually using chalk, and enable teachers to review writing more effectively in classes of over 100 students. Classrooms are also given wall clocks to help teachers keep track of time during lessons, and the program supports teacher-parent meetings once per term.[5]

2.3 The Reduced-Cost Version

Mango Tree's goal was to create the highest-quality and most-effective literacy program possible. However, 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 materials development and the NULP's broader community activities, the program costs $19.88 per student (Table 1).[6] This is more than twice the average intervention covered in McEwan (2015), and is more expensive than 94% of the 16 studies in the McEwan sample with cost data.

[5] Mango Tree also promotes local-language literacy more broadly in the community. We are unable to quantify how this contributes to the NULP's impacts but because all three study arms are exposed.

[6] Costs are calculated on a per-student basis. We use Mango Tree program expenditures from 2013 for all items except the time costs of teachers and CCTs, which we estimate from survey data ($5.74 per day) and the wall clocks, which we estimate from local markets.

Mango Tree therefore created a modified, reduced-cost version of the NULP.

There are three main differences between the full- and reduced-cost versions of the NULP (Table 1 and Appendix Table 1). The first is the use of a cascade model of training and support, rather than working directly with teachers. This approach involves Mango Tree staff directly training the government CCTs, who were tasked with conducting teacher trainings and support visits themselves. CCTs were provided with all of the NULP training materials as well as instructional videos (and solar DVD players) to show to teachers at in-service training sessions in their local communities.[7] The second difference between the full- and reduced-cost versions of the NULP is that schools in the reduced-cost version received fewer support visits than those in the full-cost version: two visits per term (from the CCTs only) instead of five (two from CCTs and three from Mango Tree staff). In both program versions, CCTs were given financial resources to make school visits and to hold training sessions. The third difference between the two versions is that classrooms in the reduced-cost version were not provided slates and wall clocks, which were seen as less-essential inputs for the program.

In all, the modifications to the full-cost program reduced the program's cost by 64%, to $7.14 per student. To further understand the differences between the two program versions, we use a set of indicators developed by Arancibia et al. (2016) to characterize in-service teacher-training programs (Appendix Table 1). Out of 51 total indicators, three (5.9 percent) differ across the two versions of the NULP. The two program variants are similar in relative terms as well as

[7] CCTs trained and supported teachers using the same tools in both the full- and reduced-cost versions of the program. Because the intervention was randomized at the school level rather than at the CCT level, spillovers are possible, although we believe this is unlikely. CCTs created separate work plans for schools in the different study arm and received no financial resources for control schools.

absolute terms. Arancibia et al. (2016) use their instrument to code 26 in-service training programs, including the two versions of the NULP. Across all pairwise comparisons (a total of 325 pairs), we compute the share of indicators that are different, excluding three indicators related to sample size. On average, pairs of programs differ on 53% of all indicators. The difference between the two NULP variants is the smallest in their dataset.[8] Mango Tree records of program implementation and delivery of the two program versions show no evidence of systematic differences in non-compliance with the program across full- and reduced-cost program versions.[9]

3 Research Design

3.1 Sample and Randomization

The study was conducted in 76 first-grade classrooms in 38 government schools across five

[8] Almost all of the training programs in Arancibia et al. (2016) met over multiple days (91 percent), with an average of almost 60 hours of training spread over an average of 9 weeks. Half of the programs followed a cascade model and the majority (73 percent) include follow-up support visits, with an average of 5.8. The majority tend to be smaller-scale with approximately 700 teachers receiving training in 60 schools. The most common trainer profiles are: Expert – university professors or graduate degree in education (coded as 2), and Local government official (4), at 33% each; the full-cost NULP was coded as Primary or secondary teachers on this indicator (1), while the reduced-cost version was coded as Local government official.

[9] Mango Tree staff drafted detailed weekly work plans and activity reports noting when any program deviations were identified. For example, meeting minutes from mid-2013, explicitly discuss the guidelines and procedures for CCTs to separately manage full- and reduced-cost program schools. The report describes procedures not being followed (e.g., a CCT not conducting all days of training) and next steps. The open communication about, and monitoring of, the NULP provide some evidence of the care with which the program was delivered.

Coordinating Centres in the Lango sub-region. Schools were eligible for the study if they met 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.[10]

Schools were assigned to one of three study arms via public lottery: control, full-cost program, and reduced-cost program, in late December 2012. Prior to the lottery, schools were grouped into stratification cells – three schools in each cell – by the researchers, based on the schools' Coordinating Centre, first-grade enrollment, and distance to the Coordinating Centre headquarters. Representatives from each school within a stratification cell drew tokens indicating treatment status from an urn.

After the second week of the 2013 academic year, enumerators collected student enrollment rosters from each school to generate an ordered list of randomly-selected students, stratified by classroom and gender. The first 25 students on the list in each of the two classrooms in a school who were present on the day enumerators conducted baseline exams were selected into the sample. These 1,900 first-grade students comprise our baseline sample.

3.2 Learning Outcomes

We assess student learning with exams administered at the beginning and end of the school year: baseline tests were conducted in the third and fourth week of the school year and endline

[10] The criteria were: a) two first-grade classrooms and teachers; b) desks and lockable cabinets for each classroom; c) a student-teacher ratio no greater than 135 during 2012 in grades one to three; d) located less than 20 km from the Coordinating Centre headquarters; e) accessible by road year round; and f) a head teacher regarded as "engaged" by the CCT. Schools also could not have previously received Mango Tree support. Head teachers were asked to assign the two best teachers in their school to their two first-grade classrooms and sign a contract with Mango Tree outlining the guidelines for study participation (both prior to treatment assignment). Schools that did not adhere to the contracts lost Mango Tree support in previous years while the program was piloted.

tests were conducted during the last two weeks of the school year. Examiners were hired and trained specifically for the testing process, were not otherwise affiliated with Mango Tree, and were blinded to the study arm assignments of the schools they visited. Exams were designed to explicitly test first-graders' basic and advanced reading and writing in Leblango.

Reading Leblango. We measure reading skills using the Early Grade Reading Assessment (EGRA). The EGRA is an internationally-recognized exam designed to serve as an "assessment of the first steps students take in learning to read" (RTI International, 2009). We use a version of the EGRA adapted to Leblango for use in Uganda by RTI (Piper 2010). The exam covers six components of reading: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The first four components involve identifying letters, sounds, and real and invented words. The last two components involve reading a passage aloud and answering comprehension questions about it.[11]

Writing Leblango. To capture students' ability to write, we use a writing assessment designed by Mango Tree. Writing tests were conducted in a group. Students were first asked to write their African surname and English given name, which were each scored separately in spelling and capitalization. Students were then asked to write about what they like to do with their friends; this was scored in seven categories: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation.[12] Each writing concept was scored on a 5-point scale.

Combined Exam Score Indices. The reading and writing exams modules differ in their number of questions and some are scored based on a student's speed while others are untimed. We

[11] Another advantage of the EGRA is that it is conceptually related to other standardized tests like PISA (Dubeck and Gove 2015), so EGRA scores can be converted into equivalents on other tests (Angrist et al. 2019).

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

present program effects on each module separately, as well as on combined outcome indices constructed using principal components analysis (PCA) to measure overall reading and writing performance. We normalize the index by dividing by the endline control-group standard deviation.[13]

3.3 Longitudinal Sample

Of the 1,900 students in our baseline sample, 78% were tested at the endline. The longitudinal sample of 1,481 students comprises our main analytical sample. Appendix Table 2 presents baseline summary statistics across each study arm, among the baseline sample, longitudinal sample, and students lost to follow-up. The baseline sample is balanced in terms of demographics and test scores, and student characteristics do not systematically correlate with attrition across study arms. Appendix Table 3 shows the predictors of attrition by baseline student characteristics, separately by study arm and pooled across all three arms. The predictors of attrition differ slightly by study arm but the differences are not statistically significant.

3.4 Empirical Methods

Regression Model

We estimate the effects of the NULP on each reading and writing test component

[13] 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 in Black and Smith (2006). Our results are robust to an alternative index that takes the unweighted average of the normalized exam components, as in Kling et al. (2007); the PCA index relates test score gains due to the treatment to the control group's progress over the year. While there are official guidelines for scoring individual sections of the EGRA (RTI International, 2009) there is no defined system for combining the scores; other papers have also constructed overall EGRA scores (Aker and Ksoll, 2019).

separately, and on overall reading and writing performance using the PCA index. Our empirical strategy relies on the random assignment of schools to the three study arms for identification. We run regressions of the form:

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

Here i indexes students and s indexes schools. y_{is} is a student's outcome at endline — typically his or her score on a particular exam or exam component. $FullCost_s$ and $ReducedCost_s$ are indicators for the school being assigned to the full- or reduced-cost versions of the program. ϵ_{is} is a mean-zero error term. β_1 and β_2 are our estimates of the effects of the full- and reduced-cost programs. We control for a vector of indicators for lottery stratification cells $\boldsymbol{L}_s$ to consistently estimate treatment effects and improve the precision of our estimates (Bruhn and McKenzie 2009). Our preferred specifications also control for the baseline value of the outcome variable, $y_{is}^{baseline}$, as specified in our pre-analysis plan.[14] To account for the fact that the treatment was randomized at the school level, we report heteroskedasticity-robust standard errors clustered by school. We present additional estimates for robustness in the Appendix: results without baseline controls, and, although we have no evidence of systematic differences in attrition across study arms, Lee bounds (Lee 2019).

Statistical Power

While we have a relatively small sample of schools, we had reason to be confident that the evaluation would be well-powered. Mango Tree had conducted Leblango EGRA exams at the beginning and end of 2010, 2011, and 2012 in program pilot schools. In 2012, Mango Tree internal evaluators also choose a sample of "comparable" non-program schools that they viewed as similar

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

to their program schools. Using those data, which come from a totally separate set of schools, we estimated an effect of 1.6 on letter name recognition. Our power calculations, specified in our pre-analysis plan, indicated that our minimum detectable effects (MDE) would be 0.33 SDs (80% power, 0.20 ICC, comparing 12 schools to another 12 schools).[15] We can also conduct post-hoc power analyses following Ioannidis et al. (2017), by using our estimated standard error to determine the MDE.[16] The MDE for 80% power is 2.8 times the standard error, or 0.38 SDs for the effect of the full-cost program on overall reading, 0.47 SDs for letter name knowledge, and 0.40 SDs for overall writing, which are well below our separate estimates of the effect of the NULP from Mango Tree's historical data.[17]

[15] Our initial calculations assumed 145 students per school, but our final sample averages just 38 students per school. We estimated a partial R-squared for past test scores of 0.7 based on the year-on-year predictive power of test scores for older students; our actual R-squared is just 0.04 because most students initially cannot read at all. The observed ICC in our data is 0.16. If we use these values instead, our MDE at 80% power is 0.50 SDs with 80% power.

[16] Post-hoc power calculations that use the estimated treatment effect are subject to type-M error and tend to show that any study with a statistically-significant treatment effect is well-powered (Gelman and Carlin 2014). McKenzie and Ozier (2019) show that using the estimated standard error to construct an MDE does not have the same issue: data generated under a DGP with a given true MDE will have an estimated MDE that is close to the true value, even if only datasets with statistically-significant treatment effects are used.

[17] Standard power calculators do not correct for issues arising from having a small number of clusters, and our corrections for our small numbers of clusters produce only *p*-values and not standard errors. As a substitute, following Ioannidis et al. (2017), we can take the half-width of the 99.5% confidence interval as an estimate of the MDE at 80% power. This is the same cutoff that is selected using the 2.8 standard error rule. When we do this using the `boottest` command (Roodman et al. 2019), we find MDEs at 80% power of 0.64 SDs for overall reading, 0.79 SDs for letter

Hypothesis Testing

We conduct all hypothesis tests in this paper using randomization inference, following Athey and Imbens (2017). This approach approximates the exact *p*-value for our observed treatment effects under the sharp null hypothesis that the treatment effect is exactly zero for all units in our sample. It also addresses the issue that cluster-robust standard errors can be too small if the number of clusters is low (Cameron et al. 2008). The typical cutoff is 50 clusters; our study has just 38.

The randomization inference procedure consists of running simulated versions of the lottery that was used to assign schools to study arms. Within each stratification cell, we randomly re-assign schools to study arms and then estimate the treatment effects for these simulated assignments using equation (1). Repeating this 1000 times gives us the distribution of treatment effects that we would expect under the null hypothesis of zero average effect, where any evident treatment effects are simply due to chance. We modify the approach of Heß (2017) to account for the multiple treatment groups in our study. For each regression, we conduct three hypothesis tests – a comparison of full-cost with control, a comparison of reduced-cost with control, and a comparison of the two treatments with each other – by permuting only the two study arms in question. All the reported *p*-values and indications of statistical significance in this paper are based on this randomization inference procedure. We also show wild cluster bootstrap *p*-values for our main results in the Appendix (Cameron et al. 2008; Roodman et al. 2019).

We use two complementary methods to correct our *p*-values for multiple comparisons. The

name knowledge, and 0.62 SDs for overall writing. The overall reading and writing MDEs are comparable to those found in Gove et al. (2017) and Piper et al. (2018), while the letter name knowledge MDE is less than half the size of the effect we estimated using non-experimental methods on a separate sample of schools.

overall reading and writing PCA-based indices avoid multiple comparisons and increase our statistical power (Kling et al. 2007). When we analyze treatment effects on individual test component or analysis using classroom variables, we report *q*-values that control for the false discovery rate using the step-up method of Benjamini and Yekutieli (2001).[18]

4 Program Effects on Learning Outcomes

4.1 Program Effects on Reading

The impacts of the two versions of the NULP on EGRA scores, estimated using equation (1), are shown in Table 2.[19] The full-cost version of the program increases letter-name knowledge by 1.01 SDs, and also has strong effects on the other five EGRA components; four of those five estimates are significant at the 0.05 level. Turning to the combined reading score index in Column 1, the full-cost program shows gains of 0.64SD, confirming that the large effect of the program is not merely an artifact of focusing on knowledge of letter names. Our estimates for the full-cost program are quite precise: we can reject test score gains smaller than 0.37SD at the 0.05 level. Lee bounds that account for attrition are also fairly tight. Our lower bound estimate for the full-cost program effect on overall the EGRA index is 0.558 and significant at the 0.01 level (Appendix

[18] This adjustment depends on the total number of comparisons as well as each *p*-value, and so requires a decision about which *p*-values are included. We include all outcomes for a given domain (e.g. all reading subtests, or all student behaviors during writing class – including variables not reported in our results tables). We pool all *p*-values across the two treatment groups. We adjust the *p*-values for the differences between the two treatment groups separately, because those tests are highly correlated with the tests for our main treatment effects. No adjustment is applied to the PCA indices summarizing our main effects on reading and writing.

[19] The estimated effects on reading are virtually unchanged in magnitude or significance when we omit the baseline exam score controls (Appendix Table 4) or when we use wild cluster bootstrap *p*-values (Appendix Table 5).

Table 6).

In contrast to the full-cost program's effect, the effect of the reduced-cost program on the EGRA index is just 0.13SD and is statistically indistinguishable from zero. The reduced-cost program improves letter-name knowledge by 0.41SD, which while still meaningful, is less than half that of the full-cost version, and is not statistically significant (p=0.106). The difference between the effects in the full- and reduced-cost program is 0.61 SDs and is statistically significant at the 0.01 level. The reduced-cost program has no statistically-significant effects on the other EGRA components, and the point estimates are all very close to zero. The Lee bounds for the reduced-cost program effects tell a similar story to our main point estimates (Appendix Table 6). The upper bound on the EGRA index and all reading components are positive and statistically significant; the lower bound estimates are insignificant and close to zero for all components except letter names. Those estimates suggest effects that are between 36 and 62 percent lower than the full-cost program upper bounds.

4.2 Program Effects on Writing

Columns 2 and 3 of Table 3 show that the full-cost version of the program has large effects on students' ability to write their first and last names, with gains of 0.922 and 1.312 SDs. The full-cost program also has positive, although statistically-insignificant, effects on students' ability to write or draw a short story (Columns 4 to 10). Altogether, the combined writing score rises by 0.45SDs, which is statistically significant at the 0.1 level (Column 1).

The reduced-cost program also has large effects on increasing students' ability to write their first and last names, although the effect is about 50% smaller than that of the full-cost program. In contrast, the reduced-cost program has uniformly negative effects on story writing, with the negative effects on Voice and Presentation reaching significance at the 0.05 level. The

combined writing score falls by 0.16 SDs, although this drop is not statistically significant. The gap between the effects of the two program variants is statistically significant for every measure of writing performance ($p<0.05$) and quantitatively large.[20]

The estimates using Lee bounds reveal a similar story (Appendix Table 10). For the full-cost program estimates, the bounds are quite tight, with the upper and lower bounds showing distinctly positive effects. In contrast, the reduced-cost program's effects on the various components of story writing are all negative even at the upper bound, and the lower bounds estimates are negative, large and statistically significant.

4.3 Cost-effectiveness

The large effects of the program naturally raise the question of cost. To compare the cost-effectiveness of the two versions of the program, we present the cost per student of each program version, as well as the cost per 0.2 SD gain and the SD gain per dollar spent for three different measures of the program's effects (Table 4). We also present results using our Lee bound estimates, which provide similar conclusions.

Using the estimated program effects on the most-basic reading skill, letter-name knowledge, the two versions are relatively comparable, with results slightly favoring the reduced-cost program. The reduced-cost version increases letter name knowledge by 0.057 SDs for each

[20] The writing test results are essentially unchanged in magnitude and significance if we omit the baseline exam score controls (Appendix Table 7), or estimate wild cluster bootstrap *p*-values (Appendix Table 8). The *p*-value for the effect of the full-cost program rises above 0.1 for the former change and falls below 0.05 for the latter; the differences in the point estimates and standard errors are both quite small. One of the 12 control schools mistakenly completed the writing test in English instead of Leblango. Our main results include this school, with the test marked in English. Our results are robust to dropping the stratification cell for this school from our sample (Appendix Table 9).

dollar spent, compared to 0.051 SDs for the full-cost program. The full-cost program is slightly more costly per student per learning gain, costing an extra 41 cents per student to raise letter name knowledge by 0.2 SDs.

Assessing cost-effectiveness based on overall reading skills reverses our conclusions. The full-cost version yields almost twice the gains in SDs per dollar compared to the reduced-cost version: 0.032 SDs vs. 0.018 SDs. Similarly, the cost per 0.2 SD increase in reading is $6.23 in the full-cost program and $11.08 in the reduced-cost version. Cost-effectiveness estimates from the combined writing score index show an even starker pattern: because the reduced-cost version of the program reduces writing performance, the cost per 0.2 SD gain from that version of the program is undefined. Instead, each dollar spent on the reduced-cost version of the program *decreases* writing performance by 0.022 SDs.

5 Mechanisms

Both the full- and reduced-cost programs introduced a set of inputs meant to support teachers and increase student learning. The full-cost version of the NULP has substantial benefits for pupil literacy across all metrics of reading and writing. In contrast, the reduced-cost version seems to achieve gains on only the most basic outcomes, letter recognition and name writing, with no gains in other areas and statistically-significant losses on some more-advanced writing skills. How does a small modification of a highly productive education program lead to negative effects for some learning outcomes? The available evidence, discussed above, suggests it is unlikely that the inputs in the reduced-cost program were simply not adequately delivered. Even so, we would not *a priori* expect declines in learning outcomes as a result of providing additional educational inputs. Because the two variants of the NULP were randomly allocated as complete packages, we cannot causally separate the effects of each individual input. Instead, we sketch a conceptual

framework to provide insight into how the reduced-cost program might have backfired. We use this framework to guide our empirical exploration of the mechanisms behind our results.

5.1 Conceptual Framework

Consider an education production function that allows for multiple inputs and multiple outcomes. Following Brown and Saks (1981, 1986) and Pritchett and Filmer (1999), teachers produce multiple student learning outcomes measured by test scores. Student learning may differ across subjects (e.g. literacy and math), learning domains (e.g. reading and writing) or skill level (e.g. advanced vs. basic). Teachers maximize utility, *U*, which is a function of student learning y_s in subject or domain *s* where $s = \{1, \dots, N\}$, and other teacher outputs, y_m.

$$U = g(y_1, \dots, y_N, y_m)$$

U has positive and diminishing marginal utility in all its arguments. There is a production function, f_s, for each subject. Learning levels y_s are determined by 1) how much of each of input is applied to the particular subject, and 2) the effectiveness of each input, which can also vary by subject or subject domain.

$$y_s = f_s(x_{s1}, \dots, x_{sJ})$$

where x_{sj} is the amount of *j* input applied to subject *s*. Inputs can be materials such as slates or books, but also include time spent teaching, and student, school, and teacher characteristics. Assume that all inputs x_{sj} (weakly) positively affect learning, such that $f_{x_{sj}} \geq 0$ for all *j*, where $f_{s,x_{sj}}$ is the marginal product of input x in producing output y_s.

The NULP could affect learning outcomes in one of two ways: either by providing new inputs or changing the productivity of inputs. These changes can in turn cause additional changes in inputs due to optimizing behavior by teachers as well as interactions between inputs. Since the marginal products of all inputs are weakly positive (by assumption), the *direct* effect of adding

inputs on test scores is always to (weakly) raise learning outcomes. However, with multiple outcomes, the *net* effect of a program like the NULP on any given learning output is ambiguous. We categorize the potential ways in which an intervention could backfire on certain outcomes into three mechanisms.

A. Substitution effects due to differential productivity enhancements. Teachers may re-optimize the allocation of inputs in response to productivity enhancements caused by the program. This is conceptually similar to income and substitution effects in consumer theory. Improving the productivity of some inputs effectively lowers the "price" of that output. If the "price" of producing reading falls by more than the "price" of producing writing, then the substitution effect would cause teachers to invest less in writing. In this case, writing test scores will only improve if the "income effect" of the productivity enhancement – which makes higher achievement in both subjects attainable – is larger than the substitution effect. In the case of the reduced-cost NULP, the productivity gains for advanced writing skills may have be sufficiently smaller than the productivity gains in other literacy outcomes and induced a net negative effect.

B. Substitution effects due to missing complementary inputs. Re-optimization may also occur as a result of the program providing inputs that are technical complements to one another or to existing inputs, that is, $\partial^2 f_{s,x_{sj}} / \partial x_{sj} \partial x_{sk} > 0$. This is conceptually similar to mechanism A, but the change in the productivity of an input comes from inputs provided by the program. This will lower the effective "price" of some outputs. For the reduced-cost NULP, one possibility is that a complementary input into the production of advanced writing skills (slates, for example), was omitted. This would make the decline in the "price" of producing those skills smaller or even zero, creating a substitution effect away from advanced writing and toward other skills.

C. Negative effects on input productivity. Third, the program may directly reduce the

productivity of some inputs for certain outcomes. When teachers are fundamentally re-trained, they may initially perform worse before eventual major improvements; this is also known as a "J-curve" (Jellison 2010). For example, new teaching methods may require practice; without the additional support provided in the full-cost NULP, reduced-cost NULP teachers may not have gotten that practice. They would therefore never reach the upward part of the curve for advanced writing skills.[21] If this effect is present, it will then be compounded by the same kind of substitution effect seen in mechanisms A and B. [22]

5.2 Identifying Mechanisms through Classroom Observation Data

To investigate what drives the difference in effectiveness across the full- and reduced-cost programs, and look for evidence consistent with the three potential mechanisms listed above, we use data from a set of detailed classroom observations. Enumerators collected classroom observations three times during the school year: once during term two, and twice during term three. Each first-grade classroom was observed during two 30-minute literacy lessons per visit, using the survey instrument in Appendix Figure 1.[23] Literacy lessons were divided into three 10-minute blocks of time. For each block, the enumerator indicated whether the teacher and students engaged

[21] Similarly, students have been found to perform better if they first make mistakes on harder tests (Hays et al. 2012).

[22] A possible fourth mechanism is that the program introduced technically competitive inputs, which would have similar implications to mechanism C.

[23] There are 72 distinct teachers in the data, and the median teacher has 18 observation blocks (six classes with three blocks each). The average number of observation blocks is 16.7, and this does not differ significantly across study arms. Our data also includes 85 observation blocks where we cannot assign a teacher ID, but we do know the school. We drop those observations from analyses that require linking the classroom observation data to student test scores.

in a range of pre-determined actions in three categories: reading, writing, and speaking/listening. Enumerators indicated the number of minutes spent on each category, the share of students participating in the activity, and the materials used. They then indicated whether they saw students do various actions and whether English or Leblango was used.[24]

We measure the use of educational materials, the share of time spent on reading and writing activities, and the share of time spent using the local language during a 30-minute lesson. We also examine the specific materials used and focus of activities (e.g., sounds, sentences, or words), separately for lessons involving reading or writing activities.

5.3 Allocation of Inputs: Materials and Time on Task

Econometric Strategy

To measure the impact of the program on input allocation, we estimate the reduced-form effects of the two program variants on the materials used and time allocation during literacy lessons. We collapse the classroom observations to the level of a 30-minute lesson and estimate:[25]

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

[24] Classroom observations can be strong predictors of student achievement in both developed countries (Kane and Staiger 2012) and the developing world (Araujo et al. 2016). Our instrument focuses on objective behaviors, similar to the Stallings tool; the CLASS tool used by Araujo et al. focuses on subjective assessments of teaching quality. The CLASS and Stallings produce measures that are well-correlated (Bruns et al. 2016).

[25] The results are substantively similar using ten-minute blocks as our units of observation. For our average classroom observation measure, the lesson-level ICC is 0.232, 77% of the variance is within-lesson, and 23% across-lesson.

where *s* indexes schools, *c* indexes classrooms, *r* indexes the round of the visit, and *l* indexes the lesson being observed. In addition to the variables that appear in equation (1), equation (2) adds vectors of indicators for each observation round ($\boldsymbol{R}_r \in \{1,2,3\}$), enumerator ($\boldsymbol{E}_{rcs}$), and the day of week of the observation ($\boldsymbol{D}_{lrcs}$). We also control for the number of observation blocks in the lesson, B_{lrcs} , because some lessons are shorter or longer than 30 minutes. ϵ_{lrcs} is a mean-zero error term. We cluster the standard errors by school. Regressions are weighted by the share of time spent on reading for reading activities, and the percent spent on writing for writing activities.[26]

Effects on Material Inputs

Table 5 presents the effects of each of the program versions on the use of materials during reading and writing activities. First, note that the control group hardly uses primers or readers at all – just 3% of the time for primers and 6% for readers, likely reflecting the low availability of those materials under the status quo. Students in the full-cost program are 16 and 6 percentage points more likely to spend time reading from primers and readers respectively, materials that the NULP provides to classrooms. The effect on primers is statistically significant at the 0.05 level. Although full- and reduced-cost classrooms both received the same primers and readers, we see a smaller effect on material use in reduced-cost classrooms; however, the differences between the two programs are not statistically significant (Columns 1 and 2).

For writing, we also see large differences in the use of materials across the two program versions. Full-cost program students are much more likely to practice writing on slates, which substitutes for writing on paper (Columns 4 and 5). In contrast, reduced-cost program students spend significantly more time than full-cost program students on "air-writing" – tracing out the

[26] We get qualitatively similar results (available upon request) if we instead use unweighted regressions, which treat lessons that are just 3% reading as being equally informative about reading instruction as 100%-reading lessons.

shapes of letters in the air (Column 3). Full-cost program students also spend less time copying their teacher's text, and more time writing their own (Columns 6 and 7). The latter gain is absent for the reduced-cost program students and the difference is statistically significant (p=0.012).

Effects on Time on Task

Table 6 shows the share of the lesson allocated to reading, writing, and speaking/listening, and the share of time the local language is used. Teachers in both program versions spend more time on reading and less on speaking and listening. The drop in speaking and listening time is 2.3 percentage points larger in the reduced-cost version of the program, although this difference is not statistically significant (Column 3, p=0.169). Teachers in the full-cost program actually spend slightly less time (3.2 percentage points less, p=0.218) on writing than the control group (Column 2). Considering that the treatment effects on writing in the full-cost program are larger than those in the reduced-cost program, the improvements in writing were probably not due to increased time on task. This finding is most consistent with mechanism C from our conceptual model, which is that the reduced-cost program reduced productivity for advanced writing skills.

Teachers in both versions of the program use Leblango more often in the classroom than those in the control group (Column 4). The difference between the use of the local language in the full- and reduced-cost versions of the program is just 3.2 percentage points and not statistically significant. Given the high base rate of mother tongue instruction – the control group already uses Leblango 69% of the time – and the relatively small effects on local language use, the additional use of the local language is likely not a key determinant of the NULP's effectiveness.[27]

[27] Since the NULP promoted the use of Leblango, we also test for spillovers onto English speaking proficiency. There is no evidence of a decline in English speaking ability in either treatment arm; for more open-ended questions on the English speaking test there are gains of about 0.30 SDs for the full-cost program (results available upon request).

5.4 Productivity

Returns to Time on Task

To examine how the two program variants affected the productivity of time spent reading or writing, we use the time on task estimates and the estimated gains in reading and writing scores to calculate the gains in student learning for every hour spent on reading or writing instruction. The results, in Table 7, indicate that time spent on reading is much more productive in the full-cost program than in the other two study arms. Students in the full-cost program gain 0.012 SDs on the EGRA for each hour spent on reading, as compared with 0.004 SDs per hour in the reduced-cost program and 0.003 SDs per hour in the control. In writing, students in the full-cost program gained 0.024 SDs in scores for every hour spent on writing, as opposed to 0.007 SDs for the control group and 0.011 SDs for the reduced-cost group.

Elements of Focus

The classroom observations data provide insight into exactly how teachers were able to use their time more productively. Table 8 presents the effects of the full- and reduced-cost programs on the specific elements of focus during reading and writing lessons. Reading activities are more likely to focus on sounds in both versions of the program, reflecting the phonics-based emphasis of the NULP (Column 1). The difference between the full- and reduced-cost versions is statistically insignificant but non-trivial in magnitude; the full-cost program spends over 40% more time on sounds than the reduced-cost program. There are no detectable differences in practicing letters or words across the three study arms (Columns 2 and 3). Because students in the full-cost program perform much better on these aspects of reading, the time spent on letters and word recognition may have been more productive in the full-cost schools than in the other two study arms. There is a large, statistically significant increase in focus on sentences in the full-cost program (Column 4).

There are also some important differences across the three study arms in elements of focus during writing lessons (Table 8, Columns 5-9). Students in both the full- and reduced-cost classes spend more time on name-writing (Column 9); the full-cost treatment effect is almost 40% larger than the reduced-cost effect, but the difference is not statistically significant (p=0.573). Critically, the reduced-cost group spends substantially *less* time than the control group on writing sentences (Column 8); this reduction is not statistically significant (p=0.199), but is nearly 50% of the control-group mean.[28] This result provides some evidence in favor of mechanisms A and B from our conceptual model, both of which suggest substitution away from inputs that target those skills.

To summarize patterns across all the classroom observation variables, we use factor analysis methods to reduce the dimensionality of the data. The methods and results, described in Appendix A and Appendix Tables 11-15, indicate that compared with the reduced-cost program, teachers in full-cost program schools are more active throughout the classroom, keep the entire class engaged, and do fewer mass exercises on the board.

5.5 Potential Complementarities

Using the classroom observations, we find changes in time use, the use of materials and the focus of literacy lessons. Some of these changes are consistent with the different explanations for the backfiring of the reduced-cost NULP from our conceptual framework. Mechanism B relies on inputs being strongly complementary to one another, and the reduced-cost NULP omitting one or more key complementary inputs. Because our experiment did not separately randomized inputs

[28] When we analyze the data at the level of an observation block the effect is significant at the 0.01 level.

to schools, we are unable to test for complementarities experimentally.[29] However, we can provide some evidence that complementarities may be part of the story.

Slates as Complements in Writing Instruction

The main difference in materials between the full- and reduced-cost versions of the program is that the reduced-cost version did not provide slates for students to use to practice writing. In our model, this could reduce advanced writing skills if the slates are complementary to other inputs in teaching writing. In this case, the drop in the "price" of producing writing is not as large in the reduced-cost program as it is in the full-cost version. As a result, a substitution effect could cause teachers to invest less in writing and more in reading instead. Our evidence is broadly consistent with this theory, although we are not able to separately test the importance of the slates.

Mediation Analyses

How much can changes in classroom observation variables explain the difference in the effects of the full- and reduced-cost programs? We use the sequential *g*-estimator of Acharya et al. (2016) to estimate what proportion of the treatment effect is explained by mediators – variables affected by the treatment that in turn influence the main outcome. We estimate the effects of the mediators on the outcome variable and use those estimates to remove the effects of the mediators from the outcome variable, creating a "demediated" outcome. Then we regress the demediated

[29] There has been surprisingly little research documenting complementarities in education. While non-experimental studies suggest that the estimated effectiveness of educational inputs is highly sensitive to functional form misspecifications (Figlio 1999), experimental evidence is limited. The McEwan (2015) meta-analysis of education experiments in developing countries finds that only 9% of studies have more than two study arms, making it impossible to study complementarities between inputs. Behrman et al. (2015), Gilligan et al. (2018), and Mbiti et al. (2017) find evidence in favor of complementarities while List et al. (2013) do not.

outcome on the treatment indicator to obtain the estimated effect of the treatment on the outcome, net of the changes in the mediators. Further estimation details are in Appendix B. We restrict the predictor variables to enter the estimates linearly. The mediation analysis results suggest that the changes in classroom observation mediators – when entered linearly – explain only a small fraction of the difference in the treatment effects across study arms: 2.0% for reading (1.1% for letter name recognition alone) and 3.7% for writing (Appendix Table 16).

Machine Learning

We can contrast how well linear mediators perform at predicting the difference in the full- and reduced-cost program effects with specifications that allow for complementarities in the production function. We do so by using machine-learning techniques to assess the predictive power of our classroom observation variables for endline test scores while allowing interactions and higher order terms. We use two machine-learning methods, KRLS and the LASSO; see Appendix C for details of our approach.

For reading, the KRLS estimator yields an R-squared of 0.19 and the LASSO gives an R-squared value of 0.20 (Appendix Table 17). The OLS estimates, in contrast, give an R-squared of 0.02, suggesting that the interactions and higher-order terms are important for explaining gains in reading test scores. For writing, KRLS can predict test scores much more successfully than the LASSO; the former yields an R-squared of 0.46, while the latter has an R-squared of 0.06, which is not much higher than the OLS R-squared of 0.04. The greater predictive power of KRLS for writing scores could suggest that complementarities matter more for writing than reading, since it automatically searches for higher-order terms and interactions while the LASSO does not.

We show the ten most important predictors selected by each machine-learning technique in Appendix Tables 18 (for reading) and 19 (for writing). The results are sensitive to which method

we use but the most striking pattern is consistent across both techniques: the best predictors are dominated by three-way interactions. This is consistent with the notion that complementarities between different inputs are important for learning; however, it is difficult to determine what combinations of inputs would lead to the most learning by reading these tables. One conclusion we can draw is that there appear to be across-subject spillovers (Graham and Hebert 2011, Graham et al. 2018). Writing activities show up as important predictors of reading and vice versa, and interactions between writing and reading activities are common in the list of the most-important predictors.

5.6 Overview of Evidence on Mechanisms

Combining the conceptual model with the classroom observations data provides insights into the mechanisms behind the program's results. First, our results are consistent with the idea that the benefits of the NULP follow a J-curve, with the returns initially being negative and then eventually recovering and becoming strongly positive. This view can be rationalized by assuming the program's new teaching strategies – especially for more-advanced skills – require practice, support, and feedback to implement correctly. The additional support visits provided by the full-cost program may have helped provide the teachers with the support needed to implement the teaching strategies correctly. Looking across the two study arms and the different skills measured on the student tests, we see a pattern that is consistent with teachers falling onto different points on the J-curve for different skills. For example, the full-cost program achieves strong gains in all reading skills, while the reduced-cost program may yield some gains in the most basic reading skill, letter name knowledge (0.4 SDs, p=0.106) but fairly tight zero effects on all advanced skills. In basic writing, both versions of the program show gains, while for advanced writing we see positive effects for the full-cost program and negative effects for the reduced-cost program. This

is implies that both program versions are on the positive portion of the J-curve for basic writing skills but that they are near the bottom of the curve for advanced writing skills – with the reduced-cost version being in negative territory. Consistent with this model, the productivity of time spent on writing actually falls in the reduced-cost program schools.

Second, complementarities may play an important part in the effectiveness of the program. We see several pieces of evidence for this. The negative effects of the reduced-cost version on advanced writing skills may have been due to a missing complementary input (the slates), causing teachers to substitute inputs away from writing and towards reading; we do see changes in elements of focus during writing activities across the two program versions, for example. Another possible complementary input could have been the support visits, which were more numerous and provided by more-experienced trainers in the full-cost version of the program. The absence of these visits in reduced-cost program schools could help explain the small effects on advanced reading skills in this study arm. Our machine learning results also lend support to the view that complementarities matter, as the most-important predictors were interactions between different classroom inputs and the evidence of spillovers across subjects.

6 Conclusion

In this paper, we document how the effectiveness of an intervention can be highly sensitive to very small changes in inputs, and that the specific outcome used to measure effectiveness matters immensely for determining a program's (cost) effectiveness; both of these phenomena can lead to misleading conclusions about how to improve learning. We compare two versions of an early-primary literacy program, randomly assigned to schools in northern Uganda: a full-cost version delivered by the organization that designed the program, and a reduced-cost version

delivered through a train-the-trainers approach, with some of the more-expensive inputs removed.

After one year, the full-cost version of the program leads to massive learning gains: reading improves by 0.64SDs and writing by 0.45SDs. We see gains around 1SD for the most basic skills: letter recognition and writing one's name. The reduced-cost version performs substantially worse. It improves only basic reading and writing outcomes, leaving advanced reading skills nearly unchanged and worsening students' advanced writing skills relative to the control group.

These qualitatively different outcomes arise from seemingly-minor differences in implementation and measurement details – the two program versions differ by only 6% on a standardized metric of the attributes of in-service teacher-training programs (Arancibia et al. 2016). Yet students in the reduced-cost version of the program experienced reading gains that were 80% smaller, and writing gains that were 135% smaller (that is, negative).

Using detailed classroom observation data, we show that changes in time on task during literacy lessons are unlikely to explain the results. We find some evidence, however, that differences in the use of, and productivity of, time and materials may be a crucial part of the story. We also show some suggestive evidence of complementarities between inputs in the education production function by comparing linear mediation analysis with a machine learning approach that allows for nonlinearities and interactions in classroom observation variables.

Our results are consistent with a conceptual framework in which the NULP affects multiple learning outcomes by providing inputs or altering the productivity of inputs. The backfiring of the reduced-cost version for advanced writing skills could be driven by teachers substituting effort and inputs away from activities that receive smaller boosts to productivity. In particular, these effects could have been driven by missing complementary inputs such as slates and additional support visits. These results could also be driven by the reduced-cost version causing actual declines in

teacher productivity, because teachers were on a downward-sloping part of the learning curve and never reached their full productivity potential.

Our results provide evidence that is consistent with a complex and multi-dimensional learning process, with multiple inputs, multiple outputs, and complementarities in education production. Providing additional inputs and training to teachers results in a reallocation of inputs and changes in input productivity; see for example Glewwe et al. (2004) who discuss how agents re-optimize behavioral responses to variations in educational inputs. This complexity in education production imply that the effectiveness of a program can be highly sensitive to small variations in certain inputs. The sensitivity to inputs may help explain the large variation in program effectiveness of interventions; for example, Conn (2017) finds a 95% confidence interval for effect sizes of 0.091 to 0.27 SDs for education programs in Sub-Saharan Africa.

This paper contributes to an ongoing debate about the validity of drawing inferences from experiments in economics and generalizability in randomized controlled trials. An extensive literature has criticized randomized experiments as being limited in their ability to guide policy and provide generalizable insights.[30] A growing body of research also documents that the effectiveness of social programs can be extremely sensitive to small differences in implementation, context, or measurement (Duflo 2017). Taken together, the body of evidence on "what works" using randomized trials may lack construct validity (Nadel and Pritchett 2016). This is a deeper issue than external validity: even if a program works equally well outside of the study setting, we

[30] See Deaton (2010), Allcott (2015), and Banerjee et al. (2017) on threats to external validity, Ludwig et al. (2011) on the difficulty of identifying mechanisms in experiments, and Harrison and List (2004) and Levitt and List (2007) on the relative validity of lab and field experiments. Davis et al. (2017) discuss how to study the effectiveness of a program as it will be implemented at scale.

may not be studying the same underlying object that would be implemented elsewhere.

Evidence on the sensitivity of program results to implementation details is scarce. A study by Bold et al. (2018) finds that an education program that generates statistically-significant gains in student test scores (by 0.18 SDs) when implemented by the NGO has no effect when implemented by the government. Similarly, Vivalt (2017) finds that government-implemented programs produce smaller impacts. Our results verify and extend these findings: we show that changes to the details of a program that are quantitatively small using objective indicators 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": it is not clear what happened in the government-implemented vs. NGO-implemented versions that resulted in the difference in effectiveness.

Finally, this study highlights the challenges of measurement in studying education programs. Metrics of learning vary widely across studies, and results are often compared in terms of SDs. Yet had we not measured both reading *and* writing outcomes and reported both basic and advanced skills, we would not have had a full picture of the effectiveness of the two versions of the program. Researchers (especially economists) should pay more attention to the type and administration of learning assessments.

A more-optimistic way of interpreting our findings is to focus on the fact that the full-cost NULP program produced enormous increases on students' reading and writing in grade one, after just a single year. This provides some hope that it is possible to produce substantial learning gains in the most poor, rural African schools, without offering monetary incentives or increases in wages,

and utilizing existing government teachers.[31] As for the reduced-cost NULP, the results remind us that teaching students how to read and write is not easy, especially in settings with poor working conditions and limited training and support (Evans and Yuan, 2018). Efforts to strip down programs in order to cut costs may make them less cost-effective, and could even cause them backfire for some outcomes.

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. (2019). Call Me Educated: Evidence from a Mobile Monitoring Experiment in Niger. *Economics of Education Review*, in press.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2019). *Measuring Human Capital* (Policy Research Working Paper No. 8742). World Bank.

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.

Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments*, 1, 73–140.

[31] This is in contrast to programs that recruit new teachers (Bold et al. 2018, Muralidharan and Sundararaman 2013, Duflo et al. 2015) or provide teachers with additional classroom help (Banerjee et al. 2007).

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

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, 122(3), 1235–1264.

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

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.

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. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, 168, 1–20.

Boone, P., Fazzio, I., Jandhyala, K., Jayanty, C., Jayanty, G., Johnson, S., … Zhan, Z. (2016). The Surprisingly Dire Situation of Children's Education in Rural West Africa: Results from the CREO Study in Guinea-Bissau (Comprehensive Review of Education Outcomes). In S. Edwards, S. Johnson, & D. N. Weil (Eds.), *African Successes, Volume II: Human Capital* (pp. 255–280). University of Chicago Press.

Brown, B. & Saks, D. (1981). The Microeconomics of Schooling. *Review of Research in Education*, 9, 209-254.

Brown, B. & Saks, D. (1986). Measuring the Effects of Instructional Time on Student Learning:

Evidence from the Beginning Teacher Evaluation Study. *American Journal of Education,* 94 (4), 480-500.

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., De Gregorio, S., & Taut, S. (2016). *Measures of Effective Teaching in Developing Countries*. RISE Working Paper Number 16/009.

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

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.

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2019). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, in press.

Conn, K. M. (2017). Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research*, 87(5), 863–898.

Davis, J. M. V., Guryan, J., Hallberg, K., & Ludwig, J. (2017). *The Economics of Scale-Up* (Working Paper No. 23925). National Bureau of Economic Research.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, *48*(2), 424–455.

Duflo, E., Dupas, P., & Kremer, M. (2015). School Governance, Teacher Incentives, and Pupil–

Teacher Ratios: Experimental Evidence from Kenyan Primary Schools. *Journal of Public Economics*, 123, 92–110.

Duflo, E. (2017). Richard T. Ely Lecture: The Economist as Plumber. *American Economic Review*, *107*(5), 1–26.

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

Evans, D. & Yuan, F. (2018). The Working Conditions of Teachers in Low- and Middle-Income Countries. RISE Working Paper.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

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

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A. M., & Neal, D. (2018). *Educator Incentives and Educational Triage in Rural Primary Schools* (NBER Working Paper No. 24911).

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

Glewwe, P., & Muralidharan, K. (2016). Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 5, pp. 653–743).

Glewwe, P., Ross, P. H., & Wydick, B. (2018). Developing Hope among Impoverished Children Using Child Self-Portraits to Measure Poverty Program Impacts. *Journal of Human Resources*, 53(2), 330–355.

Gove, A., Korda Poole, M., & Piper, B. (2017). Designing for Scale: Reflections on Rolling Out Reading Improvement in Kenya and Liberia. *New Directions for Child and Adolescent Development*, 2017(155), 77–95.

Graham, S., & Hebert, M. (2011) Writing to Read: A Meta-Analysis of the Impact of Writing and Writing Instruction on Reading. *Harvard Educational Review*, 81(4), 710-744.

Graham, S., Liu, X., Bartlett, B., Ng, C., Harris, K. R., Aitken, A., & Talukdar, J. (2018). Reading for Writing: A Meta-Analysis of the Impact of Reading Interventions on Writing. *Review of Educational Research*, 88(2), 243–284.

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.

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

Harrison, G. W., & List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Hays, M. J., Kornell, N., & Bjork, R. A. (2012). When and Why a Failed Test Potentiates the

Effectiveness of Subsequent Study. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, 127(605), F236–F265.

Jellison, J. M. (2010). *Managing the dynamics of change: The fastest path to creating an engaged and productive workplace*. McGraw Hill Professional.

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.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119.

Lee, David S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, 76 (3), 1071–1102,

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.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *The*

Quarterly Journal of Economics, in press.

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.

McKenzie, D., & Ozier, O. (2019). Why ex-post power using estimated effect sizes is bad, but an ex-post MDE is not. *World Bank Development Impact Blog*, May 16, 2019. Retrieved June 6, 2019, from: https://blogs.worldbank.org/impactevaluations/why-ex-post-power-using-estimated-effect-sizes-bad-ex-post-mde-not

Muralidharan, K., & Sundararaman, V. (2013). *Contract Teachers: Experimental Evidence from India* (NBER Working Paper No. 19440).

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*, *60*(4), 776–807.

Piper, B., Simmons Zuilkowski, S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development*, 106, 324–336.

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

Pritchett, L. & Filmer D. (1999). What Education Production Functions Really Show: a Positive

Theory of Education Expenditures. *Economics of Education Review,* 18, 223–239.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal: Promoting Communications on Statistics and Stata*, *19*(1), 4–60.

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 Assessme nt 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.

Townsend, W. (2018). *ELASTICREGRESS: Stata module to perform elastic net regression, lasso regression, ridge regression* (Version S458397). Boston College Department of Economics.

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

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

Table 1

NULP Components and Marginal Costs by Study Arm

	Full-cost program		Reduced-cost program	
	Amount	Cost per Student	Amount	Cost per Student
Pedagogy				
Local Language-First Instruction	Yes		Yes	
NULP Instructional Model	Yes		Yes	
Books				
Leblango Primers	1 per term per student (3 total per student)	$0.91	1 per term per student (3 total per student)	$0.91
Leblango Readers	1 per term per student (3 total per student)	$0.91	1 per term per student (3 total per student)	$0.91
Leblango Alphabet Chart	1 per classroom	$0.03		$0.03
Leblango Teacher's Guides	1 per classroom	$0.12	1 per classroom	$0.12
English Primers	1 per term per student (3 total per student)	$0.91	1 per term per student (3 total per student)	$0.91
English Teacher's Guides	1 per classroom	$0.12	1 per classroom	$0.12
Materials				
Slates	1 per student	$1.16		$0.00
Wall Clocks	1 per classroom	$0.13		$0.00
Training and Support for Teachers				
Literacy Methods Training (3-5 days, before term)	1X/term, residential, taught by MT staff	$8.82	1X/term, non-residential, taught by CCTs	$3.51
Saturday in-service training wkshps (1 Day, during each term)	2X/term, non-residential, taught by MT staff	$3.21	2X/term, non-residential, taught by CCTs	$0.62
Classroom support supervision	3X/term from MT staff, 2X/term from CCTs	$1.69	2X/term from CCTs	$0.00
Other				
Parent Meetings	1X/term	$1.86		
Take a Book Home Activity	At parent meeting, early during first term			
Total Cost		**$19.88**		**$7.14**

Notes: This table shows the components of each version of the NULP intervention and their marginal costs. The costs of developing the intervention and materials are not included as those are one-off costs that will not be repeated in the future. Monetary costs are drawn from a detailed expense workbook shared by Mango Tree Educational Enterprises Uganda. We also include time costs in the Training and Support for Teachers category. Time costs are only counted for days on which the person would not otherwise be working. Teacher and CCT time costs are priced at the average daily wage for a teacher in our sample; MT Staff costs are priced based on salary data from the organization.

Table 2

Program Impacts on Leblango Early Grade Reading Assessment Scores

(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Leblango EGRA Score Index†	(2) Letter Name Knowledge	(3) Initial Sound Recognition	(4) Familiar Word Recognition	(5) Invented Word Recognition	(6) Oral Reading Fluency	(7) Reading Comprehension
Full-cost program	0.638***	1.014***	0.647***	0.374**	0.215	0.476**	0.445**
S.E.	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.128)	(0.113)
R.I. p-value	[0.005]	[0.006]	[0.007]	[0.010]	[0.161]	[0.025]	[0.030]
q-value	--	{0.040}	{0.040}	{0.040}	{0.276}	{0.072}	{0.072}
Reduced-cost program	0.129	0.407	0.076	-0.002	0.031	0.071	0.045
S.E.	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
R.I. p-value	[0.327]	[0.106]	[0.415]	[0.994]	[0.675]	[0.444]	[0.668]
q-value	--	{0.212}	{0.592}	{0.994}	{0.736}	{0.592}	{0.736}
Number of students	1460	1476	1481	1474	1471	1467	1481
Number of schools	38	38	38	38	38	38	38
Adjusted R-squared	0.149	0.219	0.103	0.066	0.075	0.074	0.058
Difference between treatment effects	0.509**	0.607**	0.570***	0.376***	0.184	0.405**	0.400**
S.E.	(0.127)	(0.159)	(0.128)	(0.092)	(0.093)	(0.117)	(0.120)
R.I. p-value	[0.010]	[0.020]	[0.006]	[0.007]	[0.212]	[0.021]	[0.038]
q-value	--	{0.032}	{0.021}	{0.021}	{0.212}	{0.032}	{0.046}
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces. † 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), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Table 3

Program Impacts on Writing Test Scores

(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	Name-Writing		Story-Writing						
		African (Family) Name	English (Given) Name	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Full-cost program	0.449*	0.922***	1.312***	0.163	0.441	0.152	0.175	0.383	0.221	0.139
S.E.	(0.144)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
R.I. p-value	[0.064]	[0.001]	[0.001]	[0.536]	[0.173]	[0.539]	[0.466]	[0.231]	[0.385]	[0.558]
q-value	--	{0.009}	{0.009}	{0.558}	{0.283}	{0.558}	{0.558}	{0.347}	{0.495}	{0.558}
Reduced-cost program	-0.159	0.435**	0.450**	-0.274	-0.316	-0.313***	-0.262	-0.330	-0.253	-0.330***
S.E.	(0.122)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
R.I. p-value	[0.421]	[0.011]	[0.021]	[0.150]	[0.155]	[0.006]	[0.102]	[0.104]	[0.297]	[0.007]
q-value	--	{0.040}	{0.063}	{0.279}	{0.279}	{0.032}	{0.234}	{0.234}	{0.411}	{0.032}
Number of students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.352	0.240	0.236	0.174	0.304	0.177	0.200	0.302	0.164	0.171
Difference between treatment effects	0.608***	0.487**	0.861***	0.436***	0.757***	0.465***	0.437***	0.713***	0.474***	0.469***
S.E.	(0.128)	(0.135)	(0.154)	(0.148)	(0.173)	(0.118)	(0.139)	(0.174)	(0.151)	(0.115)
R.I. p-value	[0.004]	[0.029]	[0.001]	[0.005]	[0.000]	[0.003]	[0.008]	[0.001]	[0.005]	[0.003]
q-value	--	{0.029}	{0.003}	{0.006}	{0.000}	{0.005}	{0.009}	{0.003}	{0.006}	{0.005}
Raw (unadjusted) values§										
Control group mean	0.482	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control group SD	1.000	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Presentation (column 10), which was not one of the marked categories at baseline; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.† 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), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Table 4

Cost-Effectiveness Calculations

	(1)	(2)	(3)	(4)	(5)	(6)
		Full-cost			Reduced-cost	
	Main Estimate	Upper Bound	Lower Bound	Main Estimate	Upper Bound	Lower Bound
Cost per student per year	$19.88	$19.88	$19.88	$7.14	$7.14	$7.14
Letter Name Knowledge						
Effect size (SDs)	1.014	1.045	0.955	0.407	0.590	0.364
Cost per student/0.2 SDs	$3.92	$3.80	$4.16	$3.51	$2.42	$3.92
SDs per dollar	0.051	0.053	0.048	0.057	0.083	0.051
PCA EGRA Index						
Effect size (SDs)	0.638	0.642	0.558	0.129	0.282	0.108
Cost per student/0.2 SDs	$6.23	$6.19	$7.12	$11.08	$5.07	$13.23
SDs per dollar	0.032	0.032	0.028	0.018	0.039	0.015
PCA Writing Test Index						
Effect size (SDs)	0.449	0.512	0.305	-0.159	-0.09	-0.183
Cost per student/0.2 SDs	$8.85	$7.76	$13.03	N/A	N/A	N/A
SDs per dollar	0.023	0.026	0.015	-0.022	-0.013	-0.026

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. Main Estimates come from our main analyses in Tables 2 and 3. Upper Bound and Lower Bound columns show the Lee Bounds from Appendix Tables 6 and 10.

Table 5

Classroom Observations: Materials Used

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Materials Used during Reading		Materials Used during Writing				
	Primer	Reader	Air Writing	On Slate	On Paper	Copying Text from	Writing Own Text
Full-cost program	0.160***	0.058	-0.035	0.187**	-0.106*	-0.165**	0.241**
S.E.	(0.034)	(0.027)	(0.022)	(0.042)	(0.045)	(0.046)	(0.054)
R.I. p-value	[0.002]	[0.281]	[0.246]	[0.015]	[0.055]	[0.021]	[0.011]
q-value	{0.030}	{0.529}	{0.369}	{0.126}	{0.205}	{0.126}	{0.126}
Reduced-cost program	0.102**	0.039	0.041	0.008	0.023	-0.036	0.071
S.E.	(0.032)	(0.026)	(0.018)	(0.032)	(0.035)	(0.048)	(0.046)
R.I. p-value	[0.024]	[0.205]	[0.159]	[0.827]	[0.646]	[0.549]	[0.276]
q-value	{0.120}	{0.439}	{0.341}	{0.856}	{0.745}	{0.659}	{0.394}
Number of observation periods	398	398	326	326	326	326	326
Number of schools	38	38	38	38	38	38	38
Adjusted R-squared	0.108	0.288	0.025	0.228	0.248	0.202	0.282
Difference between treatment effects	0.058	0.018	-0.076***	0.179***	-0.129*	-0.128**	0.169**
S.E.	(0.033)	(0.024)	(0.017)	(0.042)	(0.052)	(0.044)	(0.046)
R.I. p-value	[0.279]	[0.662]	[0.002]	[0.000]	[0.081]	[0.032]	[0.012]
q-value	{0.600}	{0.764}	{0.015}	{0.000}	{0.203}	{0.120}	{0.060}
Control group mean	0.017	0.042	0.080	0.028	0.446	0.368	0.142
Control group SD	0.074	0.151	0.186	0.115	0.276	0.304	0.209

Notes: Sample is 398 lessons in which students do any reading and 326 lessons in which students do any writing, based on 440 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 enumerator, and the day of the week, as well as the average value of the observation period (1, 2, or 3) for the lesson, and are weighted by the share of time spent on reading (columns 1-2) or writing (columns 3-7) during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01.

Table 6

	Time on Task			
	(1)	(2)	(3)	(4)
	Share of Time:			
	Reading	Writing	Speaking and Listening	Percent in Leblango
Full-cost program	0.061**	-0.032	-0.030*	0.111*
S.E.	(0.015)	(0.018)	(0.013)	(0.036)
R.I. p-value	[0.023]	[0.218]	[0.081]	[0.062]
q-value	{0.090}	{0.374}	{0.182}	{0.320}
Reduced-cost program	0.052**	0.001	-0.053**	0.076
S.E.	(0.015)	(0.017)	(0.014)	(0.039)
R.I. p-value	[0.030]	[0.974]	[0.019]	[0.235]
q-value	{0.090}	{0.974}	{0.090}	{0.416}
Number of lessons	440	440	440	440
Number of schools	38	38	38	38
Adjusted R-squared	0.060	-0.021	0.253	0.171
Difference between treatment effects				
	0.009	-0.032	0.023	0.036
S.E.	(0.016)	(0.017)	(0.011)	(0.029)
R.I. p-value	[0.693]	[0.252]	[0.169]	[0.324]
q-value	{0.693}	{0.378}	{0.338}	{0.912}
Control group mean	0.318	0.241	0.433	0.691
Control group SD	0.188	0.208	0.183	0.298

Notes: Sample is 440 lesson observations for 38 schools. All regressions control for indicators for stratification cell, the round of the observations the enumerator, and the day of the week, as well as the average value of the observation period (1, 2, or 3) for the lesson. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-

Table 7

Productivity of Time on Task

	(1) Full-cost program	(2) Reduced-cost program	(3) Control
Total literacy class time in P1			
# of terms	3	3	3
Instruction weeks per term	12	12	12
Classes per week	10	10	10
Minutes Per class	30	30	30
Total literacy hours in P1	180	180	180
Reading			
Share of time spent on reading	0.379	0.370	0.318
Total hours spent on reading	68.2	66.6	57.2
Reading gain in P1	0.786	0.277	0.148
Reading gain per hour	0.012	0.004	0.003
Writing			
Share of time spent on writing	0.209	0.242	0.241
Total hours spent on reading	37.6	43.6	43.4
Writing gain in P1	0.917	0.309	0.468
Writing gain per hour	0.024	0.007	0.011

Notes: This table combines information on time use from Table 5 with the estimated gains in reading and writing by study arm from Tables 2 and 3 to estimate the productivity of each minute of class time during first grade.

Table 8

Classroom Observations: Elements of Focus

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Element of Focus During Reading				Element of Focus During Writing				
	Sounds	Letters	Words	Sentences	Pictures	Letters	Words	Sentences	Name
Full-cost program	0.106**	0.048	0.054	0.094*	0.107	-0.085	0.017	-0.011	0.136
S.E.	(0.020)	(0.028)	(0.034)	(0.036)	(0.052)	(0.044)	(0.051)	(0.052)	(0.055)
R.I. p-value	[0.011]	[0.301]	[0.333]	[0.050]	[0.191]	[0.213]	[0.802]	[0.872]	[0.111]
q-value	{0.083}	{0.529}	{0.529}	{0.177}	{0.357}	{0.357}	{0.856}	{0.872}	{0.256}
Reduced-cost program	0.075**	0.082	0.024	0.017	0.110*	0.040	0.126*	-0.074	0.099**
S.E.	(0.021)	(0.031)	(0.033)	(0.043)	(0.049)	(0.042)	(0.051)	(0.041)	(0.034)
R.I. p-value	[0.015]	[0.113]	[0.509]	[0.826]	[0.072]	[0.542]	[0.065]	[0.199]	[0.027]
q-value	{0.090}	{0.308}	{0.694}	{0.826}	{0.205}	{0.659}	{0.205}	{0.357}	{0.135}
Number of lessons	398	398	398	398	326	326	326	326	326
Number of schools	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.099	0.016	0.101	0.075	0.091	0.115	0.186	0.211	0.294
Difference between treatment effects	0.031	-0.034	0.030	0.077**	-0.003	-0.125*	-0.108	0.063	0.037
S.E.	(0.018)	(0.027)	(0.034)	(0.028)	(0.043)	(0.038)	(0.043)	(0.043)	(0.048)
R.I. p-value	[0.280]	[0.340]	[0.502]	[0.022]	[0.949]	[0.054]	[0.115]	[0.266]	[0.573]
q-value	{0.600}	{0.637}	{0.685}	{0.240}	{0.963}	{0.162}	{0.246}	{0.499}	{0.811}
Control group mean	0.046	0.161	0.622	0.320	0.181	0.194	0.326	0.160	0.094
Control group SD	0.132	0.237	0.310	0.320	0.241	0.285	0.274	0.251	0.220

Notes: Sample is 398 lessons in which students do any reading and 326 lessons in which students do any writing, based on 440 lesson observations for 38 schools. All regressions control for indicators for stratification cell, the round of the observations the enumerator, and the day of the week, as well as the average value of the observation period (1, 2, or 3) for the lesson, and are weighted by the share of time spent on reading (columns 1-2) or writing (columns 3-7) during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01.