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

Jason T. Kerwin and Rebecca L. Thornton*

January 29, 2020

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 potentially 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 Michigan, Johns Hopkins, Université Paris-Dauphine, Minnesota, CSAE, Wilfrid Laurier University, CIES, the ESRC-DFID Poverty Conference, and London Experimental Week for comments and suggestions. We also thank Victoria Brown, Bernadette Jerome, Benson Ocan, and the rest of the Mango Tree staff. Funding was provided by the Hewlett Foundation, ESRC-DFID, an anonymous donor, and the University of Michigan's Rackham Graduate School. All mistakes and omissions are our own. The online appendix is available here.

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 evaluated the effectiveness of a wide range of educational interventions across a variety of contexts.[1] Systematic reviews suggest enormous heterogeneity in effectiveness across studies, making it difficult to generalize from specific evaluations to inform policy (Nadel and Pritchett 2016). Some of this heterogeneity may be due to differences in context (e.g. India vs. Kenya) or the type of intervention evaluated (e.g. providing materials vs. upgrading infrastructure), but the variation is still substantial when holding the context or type of intervention fixed (Evans and Popova 2016, Vivalt 2017). Evidence on heterogeneity comes primarily from across-study comparisons, in part because most studies evaluate just a single intervention (McEwan 2015).[2] In contrast, this paper examines how program effectiveness varies within a single study—holding the context and 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 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,

[1] Evans and Popova (2016) summarize six systematic reviews of education program effectiveness in developing countries; another was released after their paper was published (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 ways to deliver in-service teacher training.

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 experiment to study the Northern Uganda Literacy Project (NULP), a mother-tongue-first early-primary literacy program developed by curriculum experts in Uganda. The NULP provides material inputs, high-quality teacher training, and support to first- to third-grade teachers. We compare 12 public primary schools that receive the program's entire array of inputs with 12 schools that were randomized to a control group.

At the end of first grade, mother-tongue letter recognition improves by 1.01 SDs; overall reading improves 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, at nearly $20 per student-year the program is costly for a developing-country program. To study how reducing costly inputs would change the program's effectiveness, we also evaluate a reduced-cost version of the NULP. This reduced-cost version involved three changes: 1) removing the most expensive material input; 2) a cascade model where training is delivered by government employees; and 3) fewer support visits to teachers. These changes reduced the per-student cost of the program by over 60 percent, while amounting to just a 6% difference on the Arancibia et al. (2016) indicators for in-service teacher-training programs.

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 (0.41 SDs), no

significant effects on more-sophisticated literacy skills (reading actual words or sentences), and small and statistically insignificant gains to overall reading (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 skills—the ability to write one's first name (0.45 SDs) and last name (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 education production, in which teachers maximize utility over multiple learning outcomes and the NULP affects learning by providing inputs and changing their productivity. The backfiring effects of the reduced-cost program on advanced writing skills can be explained through several mechanisms. First, if the intervention raises productivity in one skill more than another, teachers may substitute investments towards the second skill. 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 – a so-called "J-curve" for learning skills.

We explore the implications of this model using a rich set of classroom observations. We

[3] Chao et al. (2015) and Fryer and Holden (2012) find unanticipated negative consequences of education interventions; they, however, provide extrinsic incentives to students or teachers.

find no evidence that changes in time allocated to reading and writing is an important driver of our results. Both full- and reduced-cost program teachers spend 5-6 percent more time reading with students than control teachers, and 3-5 percent less time simply lecturing to students; there are no differences in time allocation across the two study arms. Mother-tongue instruction also does not drive the results: both program variants increase use of the local language by 8-11 percent.

We do find evidence which suggests that the full-cost program succeeded primarily through more-productive use of time and materials. We find that the full-cost program increases learning gains per hour reading by 4.5 times relative to the control group, as opposed to 1.6 times for the reduced-cost program. Similarly, the gains per hour of time spent writing are 2.2 times higher in the full-cost program than in the control group. The reduced-cost program makes writing time *less* productive, achieving just 66% of the control-group gains per hour. We can identify some of the ways time is used differently: during writing lessons, students in the full-cost program shift from writing on paper to writing on slates, and write their own text rather than copying from the board; there are no significant differences between the reduced-cost program and the control group. Both program variants increase the time spent on sounds and reading sentences, but the full-cost program effects are more than 50% larger for the former (p=0.28) and over five times larger for the latter (p=0.02).

We find that one likely mechanism for the backfiring of the reduced-cost program is a J-curve in the development of teaching skills: the productivity of time spent on writing actually falls in reduced-cost schools. There is also some evidence of a role for complementarities between inputs. Mediation analyses that using classroom behaviors as linear predictors can explain less than 4% of the difference in effectiveness between the full- and reduced-cost programs, 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. We do not, however, see the expected evidence of reductions in time invested into advanced writing skills that this mechanism would predict.

In summary, 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. Implementers and educators should think carefully about complementary inputs, and also be aware that re-training teachers incompletely or without proper support could result in worse outcomes than doing nothing at all.

1 Context and Intervention

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).

1.1 The Northern Uganda Literacy Project

The program we evaluate, the Northern Uganda Literacy Project (NULP), was a direct response to the poor learning outcomes in the Lango sub-region. It was developed by Mango Tree Educational Enterprises Uganda, a 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 Appendix Table 1.

1.2 The Full-Cost Version

Uganda's official policy is that students in grades one to three are to be taught in their local language before transitioning to all English instruction in grade four. In practice, 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 effects is sparse: most studies are about Spanish-language programs in the US (Rossell and Baker 2006). The one developing-country study we know of finds mother-tongue reading gains of 0.3-0.6SDs (Piper et al. 2016).

The NULP trains and supports teachers in literacy instruction in first grade, entirely in 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, receive additional training through the Teacher Development and Management System. The government approach follows a cascade/"train-the-trainer" model, 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 videos. Teachers undergo four intensive, residential teacher-training sessions on orthography and literacy methods, one prior to the school year and 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 the 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 detailed feedback about their teaching. CCTs are trained to provide the same type of feedback as the Mango Tree staff and use the same monitoring and assessment tools. CCTs are also given additional financial resources to make two additional support supervision visits per term.

Teachers in Uganda typically rely on call-and-repeat methods with a focus on memorizing whole words (Ssentanda 2014). In contrast, the NULP program uses a phonics-based approach, where students sound words out. The NULP model introduces content more slowly than the standard curriculum, providing time to cover 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.[4]

[4] Both the government curriculum and the NULP model involve 15 half-hour literacy lessons per week. The government lessons are reading (5 lessons), writing (5 lessons), news (3 lessons), and oral literature (2 lessons). The NULP lessons are story-reading (5 lessons), creative-writing (5 lessons), and word building (5 lessons).

Although schools receive capitation grants from the government to pay for instructional materials (e.g. books, chalk, and teachers' guides), the material resources are often inadequate. The NULP provides a set of primers (textbooks that cover the curriculum) and readers (books for reading practice). First-grade NULP classrooms receive slates for students to practice writing using chalk, enabling 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]

1.3 The Reduced-Cost Version

Mango Tree's goal was to create the highest-quality 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 development and broader community activities, the program costs $19.88 per student (Appendix Table 1). This is more than twice the average intervention with cost data from McEwan (2015). 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 (Appendix Tables 1 and 2). 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 then conduct teacher trainings and support visits themselves. CCTs were provided with all of the NULP training materials as well as instructional videos (and solar DVD

[5] Mango Tree also promotes local-language literacy within the community, across all study arms.

players) to show at in-service training sessions in their local communities.[6] The second difference is that reduced-cost version schools received fewer support visits: two visits per term (from CCTs only) instead of five (two from CCTs and three from Mango Tree staff). The third difference is that the reduced-cost version did not provide slates and wall clocks.

In all, the modifications 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 2). Out of 51 total indicators, three (5.9%) differ across the two versions of the NULP. The two program variants are similar in relative as well as 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 (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. Mango Tree records of program implementation and delivery of the two program versions show no evidence of systematic differences in non-compliance across the two versions.[7]

2 Research Design

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

[7] 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.

2.1 Sample and Randomization

The study was conducted in 76 first-grade classrooms in 38 government schools across five Coordinating Centres (CCs) in the Lango sub-region. Schools were eligible for the study if they met criteria deemed important by Mango Tree to support the NULP model (see Appendix A). Using school-level data collected in late 2012, 38 schools (out of 99) met these criteria. While we have a relatively small sample of schools, we had reason to be confident that the evaluation would be well-powered (see Appendix B for details).

Schools were assigned to one of three study arms via public lottery: control, full-cost program, and reduced-cost program, in late December 2012. The lottery was run within stratification groups of three, with schools matched on CC, first-grade enrollment, and distance to the CC headquarters.

In the second week of the 2013 school year, we collected enrollment rosters from each school and used them to generate a randomly-ordered list of students, stratified by classroom and gender. Our sample for each school is the first 50 students on the list who were present on the day of the baseline exams. These 1,900 first-grade students comprise our baseline sample.

2.2 Learning Outcomes

We assess student learning using baseline exams (administered in the third and fourth weeks of the school year) endline exams (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.

Reading Leblango. We measure reading skills using the Early Grade Reading Assessment (EGRA), an internationally-recognized exam designed to assess early reading (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: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension.

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.[8] Each writing concept was scored on a 5-point scale.

Combined Exam Score Indices. The subtests within each exam differ in their number of questions and some are scored based on a student's speed while others are untimed. We present program effects on each subtest separately, as well as on combined outcome indices constructed using principal components analysis (PCA) to measure overall reading and writing performance. We standardize the index by dividing by the endline control-group standard deviation.[9]

2.3 Longitudinal Sample

Of the 1,900 students in our baseline sample, 78% were tested at the endline. These 1,481 students comprise the longitudinal sample we use for analysis. 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). The predictors of attrition differ slightly by study arm but the differences are not statistically significant (Appendix Table 4).

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

[9] Our PCA score indices are weighted averages of the subtest scores, 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 standardized exam components, as in Kling et al. (2007).

2.4 Empirical Methods

Regression Model

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. $FullCost_s$ and $ReducedCost_s$ are indicators of being assigned to the full- or reduced-cost versions of the program. ϵ_{is} is a mean-zero error term. We control for a vector of stratification cell indicators L_s to improve precision (Bruhn and McKenzie 2009). We also control for the baseline value of the outcome variable, $y_{is}^{baseline}$, as specified in our pre-analysis plan.[10] Since the treatment was randomized at the school level, we report heteroskedasticity-robust standard errors clustered by school. In the appendix, we present additional estimates without baseline controls, and, although we have no evidence of systematic differences in attrition across study arms, Lee (2009) bounds.

Hypothesis Testing

All the reported *p*-values and indications of statistical significance in this paper are based on randomization inference (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. 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

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

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. 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 for multiple comparisons. First, the PCA-based indices for overall reading and writing avoid multiple comparisons and increase our statistical power (Kling et al. 2007). Second, we report *q*-values that control for the false discovery rate using the step-up method of Benjamini and Yekutieli (2001).[11]

3 Program Effects on Learning Outcomes

3.1 Program Effects on Reading

The impacts of the two NULP versions on EGRA scores, estimated using equation (1), are in Table 1.[12] The full-cost version of the program increases letter-name knowledge by 1.01 SDs, and has strong effects on the other EGRA components; four of the 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:

[11] We include all outcomes for a given domain and 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.

[12] The estimated effects on reading are virtually unchanged when we omit baseline exam score controls (Appendix Table 5) or use wild cluster bootstrap *p*-values (Appendix Table 6).

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 0.56 and is significant at the 0.01 level (Appendix Table 7).

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 (Appendix Table 7). The upper bounds on the EGRA index and all the subtests are positive and statistically significant; the lower-bound estimates are insignificant and close to zero for all components except letter names.

3.2 Program Effects on Writing

Columns 2 and 3 of Table 2 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 0.92 and 1.31 SDs. The full-cost program also has positive, although statistically-insignificant, effects on students' ability to write a short story (Columns 4 to 10). Altogether, the combined writing score rises by 0.45 SDs, which is statistically significant at the 0.1 level (Column 1).

The reduced-cost program also greatly increases 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.[13]

The estimates using Lee bounds reveal a similar story (Appendix Table 11). For the full-cost program estimates, the upper and lower bounds show distinctly positive effects. In contrast, the reduced-cost program's effects on story writing components 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 3). We also present results using our Lee bound estimates, reaching similar conclusions.

Using the estimated program effects on the most-basic reading skill, letter-name knowledge, the two versions are relatively comparable, with the results slightly favoring the reduced-cost program. The reduced-cost version increases letter name knowledge by 0.057 SDs for each dollar spent, compared to 0.051 SDs for the full-cost program. The full-cost program is slightly more costly per student learning gain, costing an extra 41 cents per student to raise letter name knowledge by 0.2 SDs.

[13] The writing test results are essentially unchanged in magnitude and significance if we omit the baseline exam score controls (Appendix Table 8), or estimate wild cluster bootstrap *p*-values (Appendix Table 9). Our results are also robust to dropping the stratification cell in which one school mistakenly completed the writing test in English instead of Leblango (Appendix Table 10).

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.

4 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 produced substantial benefits for pupil literacy across all metrics of reading and writing. In contrast, the reduced-cost version achieves gains only in letter recognition and name writing with no gains in other areas, and statistically-significant losses in some more-advanced writing skills. How does a small modification of a highly effective education program lead to negative effects for some learning outcomes? We would not *a priori* expect declines in learning outcomes as a result of providing additional educational inputs. The available evidence, discussed above, suggests it is unlikely that the inputs in the reduced-cost program were simply not adequately delivered. 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.

4.1 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: by providing new inputs or changing the productivity of inputs. These changes can 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 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. Improving the productivity of some inputs effectively lowers the "price" of producing the associated output. For example, if the "price" of producing reading falls by more than the "price" of producing writing, then teachers will invest less in writing unless the "income" effect of the extra resources is sufficiently large. Similarly, teachers may shift towards teaching sounds, while shifting away from writing sentences.

B. Substitution effects due to missing complementary inputs. If some inputs are technical complements to others, (i.e. $\partial^2 f_{s,x_{sj}} / \partial x_{sj} \partial x_{sk} > 0$) removing some inputs can reduce the productivity of other ones. This is conceptually similar to mechanism A, but the change in the productivity comes from inputs provided by the program. This will lower the effective "price" of some outputs. The negative effects of the reduced-cost NULP on advanced writing skills may have been due to a missing complementary input (e.g., slates), causing teachers to substitute inputs away from writing and towards reading.

C. Negative effects on input productivity. 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 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.

4.2 Identifying Mechanisms through Classroom Observation Data

To investigate what drives the difference in effectiveness across the full- and reduced-cost programs, we analyze data from a set of detailed classroom observations for evidence of

substitution of inputs (shifts in time allocation or material use), changes in the productivity of inputs, and evidence of complementarities between inputs. 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. Literacy lessons were divided into three 10-minute blocks of time.[14] For each block, the enumerator indicated whether the teacher and students engaged 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, the materials used, student actions, and whether English or Leblango was used.[15] We are interested in identifying differences in input allocation—in classroom time and the use of materials, and differences in input productivity.

4.3 Allocation of Inputs: Time on Task and Materials

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

[14] There are 72 distinct teachers in the data, and the median teacher has 18 observation blocks. The average number of observation blocks is 16.7 and does not differ significantly across study arms. We drop 85 observation blocks where we cannot assign to a specific teacher.

[15] Classroom observations are strong predictors of student learning developed countries (Kane and Staiger 2012). Araujo et al. (2016) show the CLASS tool, which focuses on subjective assessments of teaching quality, predicts learning in Ecuador. The Stallings tool, which is more similar to ours, produces measured that are well-correlated with the CLASS (Bruns et al. 2016).

lessons. We collapse the classroom observations to the level of a 30-minute lesson and estimate:[16]

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

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.[17]

Effects on Input Allocation

Table 4 Columns 1-3 show the share of the lesson allocated to reading, writing, and speaking/listening. 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.

[16] 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.

[17] We get qualitatively similar results if we use unweighted regressions, which, for example, treat lessons with 3% reading as being equally informative as 100%-reading lessons.

Columns 5-9 present the effects of each of the program versions on the use of materials during reading and writing activities. The control group uses primers just 3% of the time and readers just 6% of the time, 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 read from primers and readers (which are provided by the NULP) respectively; the former effect is significant at the 0.05 level. We see a smaller effect on reading material use in reduced-cost classrooms, but the difference from the full-cost program is not statistically significant (Columns 5 and 6).

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 8 and 9). In contrast, reduced-cost program students spend significantly more time than full-cost program students on "air-writing"—tracing out the shapes of letters in the air (Column 7).

4.4 Productivity

Returns to Time on Task

To examine how the two program variants affected the productivity of time, 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 Appendix Table 12, indicate that reading time 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 full-cost schools gained 0.024 SDs in scores for every hour spent on writing, as opposed to 0.007 SDs for reduced-cost and 0.011 SDs for control. The drop in productivity for writing in the reduced-cost group is consistent with

mechanism B. If these average productivity differences also reflect differences in marginal products, then we would expect reduced-cost teachers to substitute away from writing and toward reading relative to the control group. While we do see the expected differences in the treatment effects on reading and writing scores, we do not see change is time allocations toward reading and away from writing. If teachers lowered their investments in writing, they must have done so along another margin, and not in terms of time on task.

Elements of Focus

The classroom observations data provide insight into how teachers were able to use their time more productively. Appendix Table 13 presents the effects of the full- and reduced-cost programs on specific elements of focus during reading and writing lessons. Reading activities are more likely to focus on sounds in both program variants, reflecting the NULP's phonics-based approach (Column 1). While the difference is statistically insignificant, 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), but a large, statistically significant increase in focus on sentences in the full-cost program (Column 4). 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 are also some important differences across the three study arms in elements of focus during writing lessons (Appendix Table 13, Columns 5-9). Students in both full- and reduced-cost classes spend more time on name-writing (Column 9). Critically, the reduced-cost group spends substantially *less* time than the control group on writing sentences (Column 8), potentially substituting towards writing words (Column 7); the reduction in time on sentences is not

statistically significant (p=0.199), but is nearly 50% of the control-group mean. (Estimates at the observation block level yield an effect that is significant at the 0.01 level.) Full-cost program students spend less time copying their teacher's text, and more time writing on 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).

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 C and Appendix Tables 14-18, 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.[18]

4.5 Potential Complementarities

Using the classroom observations, we find changes in the use of materials, focus of literacy lessons, and overall productivity. These changes are consistent with mechanisms A and C 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. There are two candidates for such complementary inputs. The first is slates, which the full-cost NULP provides for students to practice writing. The reduced-cost program cut the slates; 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. A second candidate for a complementary input

[18] We can reject another potential driver of differences in productivity: the use of mother-tongue instruction. Both versions increase the use of Leblango by similar amounts (Table 4, Column 4).

is the additional support visits that are provided in the full-cost program but not the reduced-cost version. It is possible that these visits are complementary to the production of higher-level reading and writing skills; removing them could have caused teachers to substitute away from those skills and toward more-basic ones such as letter names and name-writing. As our experiment did not separately randomize inputs to schools, we are unable to test for complementarities experimentally.[19] Instead, we use mediation analysis and machine learning to provide some evidence that complementarities may be part of the story.

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 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 D. 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 19).

Machine Learning

[19] Experimental evidence on complementarities in education is limited. Behrman et al. (2015), Gilligan et al. (2018), and Mbiti et al. (2017) find evidence of complementarities while List et al. (2013) do not.

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 (Hainmueller and Hazlett 2014) and the LASSO (Friedman et al. 2010); see Appendix E 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 20). 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 21 (for reading) and 22 (for writing). The most striking pattern is consistent across techniques: the best predictors are dominated by three-way interactions. While it is difficult to determine what combinations of inputs would lead to the most learning from these tables, one conclusion is that there may be across-subject spillovers (Graham and Hebert 2011, Graham et al. 2018): writing activities show up as important predictors of reading and vice versa.

4.6 Overview of Evidence on Mechanisms

Combining the model with the classroom observations sheds light on the mechanisms behind the results. Our evidence is most consistent with the third mechanism: negative effects on

productivity (mechanism C). However, we cannot rule out substitution effects due to either relative productivity changes (mechanism A) or missing complementary inputs (mechanism B).

On mechanism A, substitution due to relative productivity changes, we see the expected productivity changes in reading and writing, and the expected changes in the scores on those tests. However, we see no evidence of changes in time allocation across reading and writing activities, as would be predicted by the substitution effect mechanism. We do see some substitution across materials, and also find changes in how a teacher spends class time across the three treatment arms, but these patterns do not readily correspond to what we would expect if the backfiring of the reduced-cost program were due to this mechanism.

Similarly, we see evidence for mechanism B: complementarities may play an important part in the effectiveness of the program. 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. 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. As with mechanism A, we do not see the expected reallocation of time across subjects that should happen if this mechanism is at play. However, the direct evidence that complementarities are important mitigates that limitation somewhat.

We also find evidence consistent with mechanism C, 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; such additional support visits were only provided in the full-cost program. 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 has fairly-tight zero effects on 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 matches a model in which both program versions are on the positive portion of the J-curve for basic writing skills but 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.

5 Conclusion

In this paper, we document how the effectiveness of an intervention can be highly sensitive to 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 provide evidence that changes in productivity of time spent during literacy lessons—driven by different use of time and materials—are likely 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.

The backfiring of the reduced-cost version for advanced writing skills could be driven by teachers substituting inputs away from activities that receive smaller productivity boosts, potentially driven by missing complementary inputs such as slates and additional support visits. The reduced-cost version may also have caused actual declines in teacher productivity if 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. 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; the effectiveness of social programs can also be extremely sensitive to small differences in implementation, context, or measurement (Duflo 2017).[20] Taken together, the 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 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

[20] 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.

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 in student learning in grade one, after just a single year. This shows 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.[21] 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 to cut costs may make them less cost-effective, and could even cause them backfire for some outcomes.

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

References

1. Acharya, Acharya, Matthew Blackwell, & Maya Sen. (2016). Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects. *American Political Science Review*, 110(3), 512.

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

3. Arancibia, Violeta, Anna Popova, & David Evans. (2016). *Training Teachers on the Job: What Works and How to Measure it* (World Bank Policy Research Working Paper No. 2848447).

4. Athey, Susan, & Imbens, Guido. (2017). The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments*, 1, 73–140.

5. Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Mukerji, Shobhini, Shotland, Marc, & Walton, Michael. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, 31(4), 73–102.

6. Banerjee, Abhijit, Shawn Cole, Esther Duflo, & Leigh Linden. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), 1235–1264.

7. Behrman, Jere, Susan Parker, Petra Todd, & Kenneth Wolpin. (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.

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

9. Black, Dan, & Jeffrey Smith. (2006). Estimating the returns to college quality with multiple

proxies for quality. *Journal of Labor Economics*, *24*(3), 701–728.

10. Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, & Justin Sandefur. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, 168, 1–20.

11. Boone, Peter, Ila Fazzio, Kameshwari Jandhyala, Chitra Jayanty, Gangadhar Jayanty, Simon Johnson, Vimala Ramachandrin, Filipa Silva, & Zhaoguo Zhan. (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.

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

13. Brown, Byron & Daniel Saks. (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.

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

15. Cameron, A. Colin, Jonah Gelbach, & Douglas Miller. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, *90*(3), 414–427.

16. Chao, Melody Manchi, Rajeev Dehejia, Anirban Mukhopadhyay, & Sujata Visaria. (2015). *Unintended Negative Consequences of Rewards for Student Attendance: Results from a Field Experiment in Indian Classrooms* (SSRN Scholarly Paper No. 2597814).

17. Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, & Stephen Taylor. (2019). How to improve

teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*, in press.

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

19. Davis, Jonathan, Jonathan Guryan, Kelly Hallberg, & Jens Ludwig. (2017). *The Economics of Scale-Up* (NBER Working Paper No. 23925).

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

21. Duflo, Esther, Pascaline Dupas, & Michael Kremer. (2015). School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools. *Journal of Public Economics*, 123, 92–110.

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

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

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

25. Friedman, Jerome, Hastie, Trevor, & Tibshirani, Rob. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

26. Fryer Jr., Roland, & Richard Holden. (2012). *Multitasking, Learning, and Incentives: A Cautionary Tale* (NBER Working Paper No. 17752).

27. Gilligan, Dan, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, & Derek Neal.

(2018). *Educator Incentives and Educational Triage in Rural Primary Schools* (NBER Working Paper No. 24911).

28. Glewwe, Paul, Michael Kremer, Sylvie Moulin, & Eric Zitzewitz. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics*, 74(1), 251–268.

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

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

31. Graham, Steve, Xinghua Liu, Brendan Bartlett, Clarence Ng, Karen Harris, Angelique Aitken, Ashley Barkel, Colin Kavanaugh & Joy Talukdar. (2018). Reading for Writing: A Meta-Analysis of the Impact of Reading Interventions on Writing. *Rev. of Educational Research*, 88(2), 243-284.

32. Hainmueller, Jens, & Chad Hazlett. (2014). Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Analysis*, *22*(2), 143–168.

33. Harrison, Glenn, & John List. (2004). Field Experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

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

35. Jellison, Jerald. (2010). *Managing the Dynamics of Change: The Fastest Path to Creating an Engaged and Productive workplace*. McGraw Hill Professional.

36. Kling, Jeffrey, Jeffrey Liebman, & Lawrence Katz. (2007). Experimental Analysis of

Neighborhood Effects. *Econometrica*, *75*(1), 83–119.

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

38. Levitt, Steven, & John List. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives*, *21*(2), 153–174.

39. List, John, Jeffrey Livingston, & Susanne Neckermann. (2013). *Harnessing Complementarities in the Education Production Function* (Working Paper).

40. Ludwig, Jens, Jeffrey Kling, & Sendhil Mullainathan. (2011). Mechanism Experiments and Policy Evaluations. *Journal of Economic Perspectives*, *25*(3), 17–38.

41. Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, & Rakesh Rajani. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. *Quarterly Journal of Economics*, in press.

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

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

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

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

46. Piper, Benjamin, Stephanie Zuilkowski, & Salome Ong'ele. (2016). Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya. *Comparative Education Review*, *60*(4), 776–807.

47. Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei, & Simon King. (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.

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

49. Pritchett, Lant & Deon Filmer. (1999). What Education Production Functions Really Show: a Positive Theory of Education Expenditures. *Economics of Education Review,* 18, 223–239.

50. Roodman, David, James MacKinnon, Morten Nielsen, & Matthew Webb. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal: Promoting Communications on Statistics and Stata*, *19*(1), 4–60.

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

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

53. Ssentanda, Medadi. (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.

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

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

56. Webley, Katy. (2006). *Mother Tongue First: Children's Right to Learn in their Own*

Languages (No. id21). Development Research Reporting Service, UK.

Table 1

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 2

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 3

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 4

Classroom Observations: Input Allocation

	Panel A: Time on Task				Panel B: Materials Used				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Share of Time:				Materials Used during Reading		Materials Used during Writing		
	Reading	Writing	Speaking and Listening	Percent in Leblango	Primer	Reader	Air Writing	On Slate	On Paper
Full-cost program	0.061**	-0.032	-0.030*	0.111*	0.160***	0.058	-0.035	0.187**	-0.106*
S.E.	(0.015)	(0.018)	(0.013)	(0.036)	(0.034)	(0.027)	(0.022)	(0.042)	(0.045)
R.I. p-value	[0.023]	[0.218]	[0.081]	[0.062]	[0.002]	[0.281]	[0.246]	[0.015]	[0.055]
q-value	{0.090}	{0.374}	{0.182}	{0.320}	{0.030}	{0.529}	{0.369}	{0.126}	{0.205}
Reduced-cost program	0.052**	0.001	-0.053**	0.076	0.102**	0.039	0.041	0.008	0.023
S.E.	(0.015)	(0.017)	(0.014)	(0.039)	(0.032)	(0.026)	(0.018)	(0.032)	(0.035)
R.I. p-value	[0.030]	[0.974]	[0.019]	[0.235]	[0.024]	[0.205]	[0.159]	[0.827]	[0.646]
q-value	{0.090}	{0.974}	{0.090}	{0.416}	{0.120}	{0.439}	{0.341}	{0.856}	{0.745}
Number of lessons	440	440	440	440	398	398	326	326	326
Number of schools	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.060	-0.021	0.253	0.171	0.108	0.288	0.025	0.228	0.248
Difference between treatment effects	0.009	-0.032	0.023	0.036	0.058	0.018	-0.076***	0.179***	-0.129*
S.E.	(0.016)	(0.017)	(0.011)	(0.029)	(0.033)	(0.024)	(0.017)	(0.042)	(0.052)
R.I. p-value	[0.693]	[0.252]	[0.169]	[0.324]	[0.279]	[0.662]	[0.002]	[0.000]	[0.081]
q-value	{0.693}	{0.378}	{0.338}	{0.912}	{0.600}	{0.764}	{0.015}	{0.000}	{0.203}
Control group mean	0.318	0.241	0.433	0.691	0.017	0.042	0.080	0.028	0.446
Control group SD	0.188	0.208	0.183	0.298	0.074	0.151	0.186	0.115	0.276

Notes: Sample is 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. Panel B weights regressionsn by the share of time spent on reading (columns 1-2) or writing (columns 3-5) 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.