MAKING THE GRADE: THE SENSITIVITY OF EDUCATION PROGRAM EFFECTIVENESS TO INPUT CHOICES AND OUTCOME MEASURES

Jason T. Kerwin and Rebecca L. Thornton*

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 0.64 SD and writing scores by 0.45 SD. A reduced-cost version instead yields statistically insignificant reading gains and some large negative effects (−0.33 SDs) 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.

I. 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 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 & Pritchett, 2016). Some of this heterogeneity may be due to differences in context (e.g., India versus Kenya) or the type of intervention evaluated (e.g., providing materials versus upgrading infrastructure), but the variation is still substantial when holding the context or type of intervention fixed (Evans & Popova, 2016; Vivaldi, 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.

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, 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 twelve public primary schools that receive the program’s entire array of inputs with twelve schools that were randomized to a control group.

At the end of first grade, mother-tongue letter recognition improves by 1.01 SD; overall reading improves by 0.64 SD. The program also improves the ability to write one’s first name by 1.31 SD, write one’s last name by 0.92 SD, and overall writing performance by 0.45 SD. 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: (a) removing the most expensive material input, (b) switching to a cascade model where training is delivered by government employees, and (c) providing fewer support visits to teachers. These changes reduced the per-student cost of the program by over 60%, while amounting to just a 6% difference on the Arancibia, Popova, and Evans (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 SD), no

Received for publication August 20, 2018. Revision accepted for publication November 13, 2019. Editor: Asim I. Khwaja.

*Kerwin: Department of Applied Economics, University of Minnesota; Thornton: Department of Economics, University of Illinois.

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.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_00911.

1Evans and Popova (2016) summarize six systematic reviews of education program effectiveness in developing countries; another was released after their paper was published (Glewwe & Muralidharan 2016).

2Notable exceptions include Bold et al. (2018) who test the effectiveness of NGO versus government program delivery and Cilliers et al. (2020) who test ways to deliver in-service teacher training.
significant effects on more sophisticated literacy skills (reading actual words or sentences), and small and statistically insignificant gains to overall reading (0.13 SD, \(p = 0.33\)). The effectiveness of the two program versions diverges 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 SD) and last name (0.44 SD). At the same time, there are large, statistically significant negative effects on the components that involved writing sentences (−0.33 SD). 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 potential mechanisms. First, if the intervention raises productivity in one skill more than another, teachers may substitute investments toward the second skill. Second, a similar pattern can occur if there are important complementarities between inputs and one is omitted.

Second, 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 find no evidence that changes in time allocated to reading and writing are an important driver of our results. Full- and reduced-cost program teachers spend 5% to 6% more time reading with students than control teachers and 3% to 5% 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% and 11%, respectively.

We do find evidence suggesting 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.

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

We show several different tests for overfitting.
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 poses 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.

B. The Full-Cost Version

Uganda’s official policy is that students in grades 1 to 3 are to be taught in their local language before transitioning to all-English instruction in grade 4. 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 United States (Rossell & Baker, 1996). The one developing country study we know of finds mother-tongue reading gains of 0.3 to 0.6 SD (Piper, Zuilkowski, & Ong’ele, 2016).

The NULP trains and supports first-grade teachers in literacy instruction 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 center 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, six in-service training workshops are held on Saturdays throughout the year. 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 to each school 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 16 of the 25 letters of the Leblango alphabet are taught in first grade, with the remainder taught in grade 2. Teachers are also provided with scripted lesson plans for each literacy lesson.5

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 on 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.6

C. 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 training, 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 players) to show at in-service training sessions

5Both the government curriculum and the NULP model involve fifteen half-hour literacy lessons per week. The government lessons are reading (five lessons), writing (five lessons), news (three lessons), and oral literature (two lessons). The NULP lessons are story reading (five lessons), creative writing (five lessons), and word building (five lessons).

6Mango Tree also promotes local-language literacy within the community across all study arms.
in their own communities. The second difference is that reduced-cost program 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 twenty-six 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 data set. 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.  

III. Research Design

A. Sample and Randomization

The study was conducted in seventy-six first-grade classrooms in thirty-eight government schools across five coordinating centers (CCs) in the Lango subregion. Schools were eligible for the study if they met criteria deemed important by Mango Tree to support the NULP model (see appendix A). Thirty-eight schools (out of 99) met these criteria using school-level data collected in late 2012. 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).

In late December 2012, schools were assigned to one of three study arms via public lottery: control, full-cost program, and reduced-cost program. 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 fifty students on the list who were present on the day of the baseline exams. These 1,900 first-grade students comprise our baseline sample.

B. Learning Outcomes

We assess student learning using baseline exams (administered in the third and fourth weeks of the school year) and 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 format, with students answering the tests individually. 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. 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.

1. 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.

2. 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.

3. Presentation was added as a scoring category for endline and was not included at baseline.

4. 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, Lieberman, and Katz (2007).
C. 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).

D. 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

\[ y_{it} = \beta_0 + \beta_1 FullCost_t + \beta_2 ReducedCost_t + L_i' \gamma + \nu_{baseline}^{i} + \epsilon_{it} \]  

(1)

Here \( i \) indexes students and \( s \) indexes schools. \( y_{it} \) is a student’s outcome at endline. \( FullCost_t \) and \( ReducedCost_t \) are indicators for being assigned to the full- or reduced-cost versions of the program. \( \epsilon_{it} \) is a mean-zero error term. We control for a vector of stratification cell indicators \( L_i \) to improve precision (Bruhn & McKenzie 2009). We also control for the baseline value of the outcome variable, \( y_{baseline}^{i} \), as specified in our pre-analysis plan.\(^{11}\) 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 & 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, Gelbach, & Miller, 2008). The typical cutoff is 50 clusters; our study has just 38. Within each stratification cell, we randomly reassign schools to study arms and then estimate the treatment effects for these simulated assignments using equation (1). Repeating this 1,000 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. 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, Liebman, & Katz, 2007). Second, we report \( q \)-values that control for the false discovery rate using the step-up method of Benjamini and Yekutieli (2001).\(^{12}\)

IV. Program Effects on Learning Outcomes

A. Program Effects on Reading

The impacts of the two NULP versions on EGRA scores, estimated using equation (1), are in table 1.\(^{13}\) 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 0.64 SD, 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 0.37 SD at the 0.05 level. Lee bounds that account for attrition are also fairly tight. Our lower-bound estimate for the full-cost program’s effect on the overall EGRA index is 0.56 SD 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 0.13 SD and is statistically indistinguishable from zero. The reduced-cost program improves letter-name knowledge by 0.41 SD, less than half the effect of the full-cost version, and it is not statistically significant (\( p = 0.11 \)). The difference between the effects in the full- and reduced-cost program is 0.61 SD 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.

\(^{11}\)https://www.socialscienceregistry.org/docs/analysisplan/36/document.

\(^{12}\)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.

\(^{13}\)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).
### 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 SD, 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 SD, 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.\footnote{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 Lebango (appendix table 10).}

The estimates using Lee bounds reveal a similar story (appendix table 11). For the full-cost program estimates, the upper and lower bounds show 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.

### C. Cost-Effectiveness

The large effects of the program naturally raise the question of costs. 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 (see table 3). We also present results using our Lee bound estimates, reaching similar conclusions.

Using the estimated treatment 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 SD for each dollar spent compared to 0.051 SD for the full-cost program. The full-cost program is slightly more costly per student learning gain: an extra 41 cents per student to raise letter name knowledge by 0.2 SD.

Assessing cost-effectiveness based on overall reading skills reverses our conclusions. The full-cost version yields almost twice the gains in SD per dollar compared to the reduced-cost version: 0.032 SD versus 0.018 SD. 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 SD.

### V. 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 achieved 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 suggests it is unlikely that the inputs in the reduced-cost program were 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.

### Table 3—Cost-Effectiveness Calculations

<table>
<thead>
<tr>
<th></th>
<th>Full-Cost</th>
<th>Reduced-Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Estimate</td>
<td>Upper Bound</td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Cost per student/year</td>
<td>$19.88</td>
<td>$19.88</td>
</tr>
<tr>
<td>Letter Name Knowledge</td>
<td>$3.92</td>
<td>$3.80</td>
</tr>
<tr>
<td>SD per dollar</td>
<td>0.051</td>
<td>0.053</td>
</tr>
<tr>
<td>PCA EGRA Index</td>
<td>0.638</td>
<td>0.642</td>
</tr>
<tr>
<td>SD per dollar</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>PCA Writing Test Index</td>
<td>$8.85</td>
<td>$7.76</td>
</tr>
<tr>
<td>SD per dollar</td>
<td>0.023</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Costs based on actual expenditures on each program variant in 2013 plus the opportunity costs of teacher and CCT time. Only incremental costs are considered, and not costs related to materials development, curriculum design, and so on. Main Estimates come from our main analyses in tables 1 and 2. Upper Bound and Lower Bound columns show the Lee Bounds from appendix tables 6 and 10.
A. 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 levels (e.g., advanced versus basic). Teachers maximize utility, \( U \), which is a function of student learning \( y_s \) in subject or domain \( s \) where \( s = \{1, \ldots, N\} \), and other teacher outputs, \( y_m \):

\[
U = g(y_1, \ldots, y_N, y_m).
\]

\( U \) 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 how much of each of input is applied to the particular subject and the effectiveness of each input, which can also vary by subject or subject domain

\[
y_s = f_s(x_{s1}, \ldots, x_{sd})
\]

where \( x_{sj} \) is the amount of input \( j \) 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 \( f_{sx_{sj}} \geq 0 \) for all \( j \), where \( f_{sx_{sj}} \) 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 by changing the productivity of inputs. These changes can cause additional changes in inputs due to optimizing behavior by teachers, as well as interactions among the 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 reoptimize 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 toward 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., \( \partial^2 f_{sx_{sj}}/\partial x_{sj} \partial x_{sk} > 0 \)) removing some inputs can reduce the productivity of others. 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 toward reading.

C—Negative effects on input productivity. The program may directly reduce the productivity of some inputs for certain outcomes. When teachers are fundamentally retrained, 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.

B. 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 2 and twice during term 3. 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. For each block, the enumerator indicated whether the teacher and students engaged in a range of predetermined 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. We are interested in identifying differences in input allocation—in classroom time and the use of materials, and differences in input productivity.

---

\[15\text{There are 72 distinct teachers in the data, and the median teacher has 18 observation blocks. The average number of observation blocks is 16.7 and does not differ significantly across study arms. We drop 85 observation blocks that we cannot assign to a specific teacher.}\]

\[16\text{Classroom observations are strong predictors of student learning developed countries (Kane & Staiger, 2012). Araujo et al. (2016) show the CLASS tool, which focuses on subjective assessments of teaching quality, predicts learning in Ecuador. The Stallings tool, which is more similar to ours, produces measured that are well-correlated with the CLASS (Bruns, De Gregorio, & Taut, 2016).}\]
C. 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 lessons. We collapse the classroom observations to the level of a 30-minute lesson and estimate:

\[
y_{ircs} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + L^*_{s} + R^*_s \delta + E^*_ircs \rho + D^*_ircs \mu + \omega B_{ircs} + \epsilon_{ircs},
\]

(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 (3) adds vectors of indicators for each observation round \( (R_s \in \{1, 2, 3\}) \), enumerator \( (E_{ircs}) \), and the day of the week \( (D_{ircs}) \). We also control for the number of observation blocks in the lesson, \( B_{ircs} \), because some lessons are shorter or longer than 30 minutes. \( \epsilon_{ircs} \) 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 share spent on writing for writing activities.

Effects on input allocation. Columns 1 to 3 in table 4 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, \( p = 0.17 \)). Teachers in the full-cost program actually spend slightly less time (3.2 percentage points less, \( p = 0.22 \)) 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.

Columns 5 to 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, whereas reduced-cost program students spend significantly more than 10 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).

D. 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 SD on the EGRA for each hour spent on reading, as compared with 0.004 SD per hour in the reduced-cost program and 0.003 SD per hour in the control group. In writing, students in full-cost schools gained 0.024 SD in scores for every hour spent on writing, as opposed to 0.007 SD for reduced-cost and 0.011 SD 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, 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 in 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 did. 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 toward writing words (column 7); the reduction in time on sentences is not statistically significant ($p = 0.20$) 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 ($p = 0.01$).

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 to 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. 19

E. 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 eliminated 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 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. Because our experiment did not separately randomize inputs to schools, we are unable to test for complementarities experimentally. 20 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

---

19 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).

g-estimator of Acharya, Blackwell, and Sen (2016) to estimate what proportion of the treatment effect is explained by mediators—variables affected by the treatment that in turn influence the main outcome. 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. 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 & Hazlett, 2014) and the Lasso (Friedman, Hastie, & Tibshirani, 2010; see appendix E for the details of our approach).

For reading, the KRLS estimator yields an $R^2$ of 0.19 and the Lasso gives an $R^2$ value of 0.20 (appendix table 20). The OLS estimates, in contrast, give an $R^2$ 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 $R^2$ of 0.46, while the latter has an $R^2$ of 0.06, not much higher than the OLS $R^2$ 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 & Hebert, 2011; Graham et al., 2018): writing activities show up as important predictors of reading and vice versa.

F. 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 toward 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 of endline test scores 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 provided only in the full-cost program. Looking across the two study arms and the different skills measured on the student tests, we see a pattern 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 ($0.4 \text{ SD, } p = 0.11$) 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 variants 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.

VI. 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 0.64 SD and writing by 0.45 SD. We see gains around 1 SD 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 the 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 have also caused actual declines in teacher productivity if teachers were on a downward-sloping part of the skill development curve and never reached their full productivity potential.

Our results provide evidence consistent with a complex and multidimensional 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 Glewwe et al., 2004, who discuss how agents reoptimize 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 SD 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).21 Taken together, the evidence on what works using randomized trials may lack construct validity (Nadel & Pritchett, 2016). This is a deeper issue than external validity: even if a program works equally well outside 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 SD) 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 versus 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 standard deviations. 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.

21 See Deaton (2010), Allcott (2015), and Banerjee et al. (2017) on threats to external validity, Ludwig, Kling, and Mullainathan (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.
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 1 after just a single year. This shows it is possible to produce substantial learning gains in the poorest rural African schools without offering monetary incentives or increases in wages, and utilizing existing government teachers.22 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 & 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.

22This contrasts with programs that recruit new teachers (Bold et al. 2018; Muralidharan & Sundararaman, 2013; Duflo, Dupas, & Kremer, 2015) or provide additional classroom help (Banerjee et al., 2007).

REFERENCES


Kane, T., and D. O. Staiger, Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains, Bill and Melinda Gates Foundation (2012).
Piper, Benjamin, “Uganda Early Grade Reading Assessment Findings Report: Literacy Acquisition and Mother Tongue” (Research Triangle Park, NC: RTI International, 2010).
Piper, Benjamin, Stephanie Zuilkowski, and Salome Ong’ele, “Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya,” *Comparative Education Review* 60 (2016), 776–807.