Making the Grade:
Understanding What Works for Teaching Literacy in Rural Uganda

Jason Kerwin, University of Michigan
Rebecca Thornton, University of Michigan

April 2015

PRELIMINARY AND INCOMPLETE – PLEASE DO NOT CITE OR CIRCULATE

This paper evaluates an early primary literacy program in Northern Uganda. Through a randomized experiment, we measure the effects of the literacy program as implemented by the organization that developed it. We compare those results to a second treatment group, which received a reduced-cost version of the program that was implemented through the government and designed to simulate how the program could be implemented at scale. The full version of the program has extremely large impacts on student learning: it improves student recognition of letter names by 1.0 SD, which is among the largest impacts ever measured in a randomized trial of an education program. The reduced-cost version improves letter-name knowledge by 0.4 SDs making it slightly more cost-effective than the full version. However, its effects on overall literacy are statistically-insignificant and it generates large negative effects on certain aspects of writing. This suggests that cost-effectiveness in improving the "headline" outcome measures emphasized by programs can come at the cost of lower performance in other areas.

1 Introduction

One of the major development successes of the past several decades has been the increased access to primary education. Primary school enrollment and completion rates have grown worldwide, and particularly in sub-Saharan Africa, which had the world's highest increase in primary school enrollment – up 42 percent from 1999 to 2006 (UNESCO, 2009). However, successes in getting students to school have not been accompanied by improvements in learning or increases in basic metrics such as literacy. Governments and policy organizations have now shifted their focus to raising the quality of education, rather than just its quantity, and translating years of education into improved learning.

A large body of research has shed light on the effectiveness of various education interventions on learning. However, the majority have shown relatively small effects. A meta-analysis of 77 randomized trials of primary education programs in developing countries found the average mean effect size was an increase in 0.14 standard deviations (McEwan 2013).

This paper evaluates a primary literacy program in rural Uganda for Primary 1 students, using a randomized experiment. The literacy program that we evaluate combines multiple educational components including a mother-tongue-first instructional approach, a revised curriculum, locally-appropriate teaching materials, extensive teacher support and training, and parent engagement. In contrast to previous studies, we find large, precisely measured effects of the program on learning: letter name knowledge, improves by 1.04 SDs of the control-group score distribution. Taking the average across an index of all six components of a standardized reading test, the effect is still 0.80 SDs.

The experiment also studies a more-scalable, lower-cost version of the program in order to help shed light on issues of scalability and cost-effectiveness. The second variant is implemented at significantly lower cost, by conducting teacher training and monitoring through the existing Coordinating Centre Tutors, government employees charged with training and supporting primary school teachers in Uganda. It also provides fewer teaching materials, in particular omitting the writing slates provided to the full-cost version of the program. This reduced-cost version of the program has smaller effects, improving letter name knowledge scores by 0.42 SDs and the index of all reading test components by just 0.15 SDs, with the latter not reaching conventional levels of statistical significance.

We examine other outcomes to shed light on the possibly mechanisms for the large effects. We find through student surveys that students increase their confidence in their ability and there is suggestive evidence that they increase their enthusiasm – although not effort – in school. We also find differences in teachers behavior in the classroom where they shifted to mother-tongue instruction and activities, and spent less time bringing students back on task.

A cost-effectiveness comparison of the two programs reveals the low-cost version to be slightly more cost-effective than the full-cost one, at 0.09 SDs of letter name knowledge per dollar as opposed to 0.07 for the full-cost variant. However, focusing on the "headline" measure of letter name knowledge hides significant drawbacks to the low-cost version of the program: the cost-effectiveness result is reversed when considering the overall reading score index, and the low-cost version of the program causes a small (but statistically-insignificant) decline in students' English speaking ability, whereas the full-cost version improves performance on the subtests of the English exam that are free-form and open-ended. Most concerningly, the low-cost program causes large and statistically-significant reductions in several aspects of writing ability – of about 0.3 SDs – relative to the control group. These reductions are despite the fact that on the writing test the "headline" measure (in this case the ability to write one's

name) once again improves. In contrast, the full-cost version of the program improves writing scores across the board, with the effects on several exam components being statistically significant.

The remainder of this paper proceeds as follows. In Section 2, we describe the details of the literacy intervention Section 3 describes the research design and Section 4 the sources of data we use. Section 5 outlines our empirical strategy. Our results are broken into two sections: the effects of the two program variants on test scores are presented in Section 6, and the effects on intermediate outcomes that shed light on the mechanisms at work are in Section 7. Section 8 concludes.

2 NULP Primary Literacy Program

2.1 Background

We evaluate a primary literacy-promotion program called the Northern Uganda Literacy Project (NULP), developed by Mango Tree Educational Enterprises Uganda.[1] Mango Tree, a private, locally-owned education company, has been operating in northern Uganda in the Lango Sub-region since 2009. Within this area there are over two million people, mostly of the Langi tribe, who speak Leblango. A civil war led by the Lord's Resistance Army from 1987-2007 had a devastating impact on the region, which to date suffers severe infrastructure shortages, extreme poverty and poor access to quality education. In addition to these challenges, the region's schools show extremely poor learning outcomes, especially in terms of literacy. An assessment of early grade reading conducted by RTI in 2009 showed that over 80 percent of students in the Lango Sub-region were nonreaders at the end of P2, meaning that they could not read a single word out of a chosen paragraph. Another assessment from November 2010 found that almost none of students in the study could recognize and read a single letter by the end of P1.

2.2 Mango Tree Model of Instruction

To address this challenge, Mango Tree began working with teachers, local language boards, and government officials in 2009, to develop an innovative new educational paradigm, the NULP. The NULP focuses on P1 to P3 students, employing a mother-tongue-first instructional approach and extensive teacher support and training. We outline the main features of the program below.

Mother-Tongue Instruction

The basis of the NULP model is mother tongue instruction, which means that children are taught in the language they grew up speaking, rather than a different language that they first encounter in school. It is common across the world, and especially in Africa, for children to enroll in school and immediately begin learning in a language that they do not understand. This other language is frequently a colonial language; English is used as the *de facto* language of instruction in primary schools throughout Uganda. Learning may happen through complete immersion, where all subjects are taught in English, or where some subjects are taught in the students' mother tongue while students are also immersed in English speaking, reading, and writing from the first day of school.

Bilingual education has numerous benefits, and parents and teachers often have strong preferences for students to learn English. However, full immersion in reading and writing a language that students

[1] Uganda's primary school system numbers the levels from P1 up to P7. P1 is the first grade level offered in government schools, and the official minimum age for enrollment is 6.

do not yet know can also have powerful drawbacks. Children often simply learn to memorize and copy words, letters, and numbers, without gaining any understanding of what they are doing or how it connects to spoken words or meaning. This works against research that finds that students learn best by building on what they already know and working from simple concepts to more complex ones. Previous research suggests that education systems that use a language unfamiliar to children in school, and simply hope that children will pick up that language, are failing (Webley et al, 2006).

Despite the common practice of immersing students in a national language for literacy class, several countries including Uganda have explicit policies mandating "mother-tongue instruction" for primary schools, which means that the primary language of instruction should be students' native language. In Uganda, this policy is not entirely enforced by schools, and teachers on not trained in local orthographies. The Mango Tree program teaches literacy in P1 entirely in the students' mother tongue. Oral English is given as a subject, but no English is written on the board or for students to read.

Teacher training and on-going support

The NULP provides extensive training and support for teachers in the program's classrooms. Mango Tree's training approach focuses on the uptake of practical and appropriate classroom skills. The first teacher training module involves a five day residential workshop on the Leblango orthography, including grammatical features and letter names and sounds. Teachers also undergo three additional intensive, residential trainings on literacy methods (both whole language and phonics approaches) during the school holidays. Teachers also participate in six Saturday in-service training workshops throughout the school year.

Teaching Materials

Mango Tree developed NULP materials continuously since 2010 in partnership with teachers and local government education officials. Mango Tree's primers and readers are small and easy to store in the classroom. Classrooms are provided with slates that allow each student to practice writing individually, and to assist the teacher to review their work effectively in classes of over 100 students with limited walking space (children can hold up their slates to show their work).

Pace and Repetition

The NULP model introduces content slowly, providing time for repetition and revision. This slower instructional pace allows for students to develop necessary pre-and early literacy skills and gives more time to prepare teachers for phonics instruction. Every teacher is also provided with teachers' guides that provide a script for each literacy lesson. Four literacy lessons are taught each day in the same order. This provides teachers, who have hugely varying and underdeveloped capacities and experiences creating effective literacy lesson plans, with easy-to-remember steps that become routine over time.

Parent and Community Engagement

Part of the NULP model involves engaging with parents and the local community to communicate the benefits of mother tongue instruction. Three parent meetings are held each year to discuss language of instruction, as well as how to assess and support children's learning and literacy development at home. This involves parent training on how to interpret their child's literacy report card, and how to use a simple reading assessment tool at home. These tools are developed by the program; the assessment allows parents to know their child's performance in key literacy skills. The study will collect

data from parents that will allow assessment of parent engagement with schools and with their children. Our research design also allows us to assess intra-household impacts on learning among siblings of children in the program.

2.3 Lower-Cost Model of Instruction

To reach scale, an educational program must be both cost-effective, and sustainable in the rural African setting. In terms of cost, the most expensive inputs of the Mango Tree program are the materials (readers, teacher manuals and slates) and teacher training and support. In addition to measuring the effect of the full Mango Tree program, we also tested the mode of delivery of the program with a scaled down model of the program.

The lower-cost model of instruction was explicitly designed to realistically demonstrate how the program might be scaled up for adoption by a larger set of schools. This involved cutting the per-school cost of implementation in two ways. First, the set of materials provided, and the intensity and cost of the trainings and support provided, was reduced relative to the standard Mango Tree Program. Second, the trainings and support for teachers were provided through the employees of the Ministry of Education and Sports (MoES) who are ordinarily tasked with training and supervising teachers in Ugandan primary schools. These employees are known as Coordinating Centre Tutors (CCTs), because each one manages a set of schools near an administrative office known as a Coordinating Centre (CC). We refer to this low-cost version of the program as the CCT Program.

In this study we compare the Standard Mango Tree Program and the Government Administered Program to a control group. The details of the inputs of each program are found in Table 1 and Appendix A.

3 Research Design

In this section, we describe the research design that underlies this study. Figure 1 illustrates the selection and randomization.

3.1 Sample

Selection of Schools

The evaluation was conducted among 38 eligible schools located in the five Coordinating Centres with existing Mango Tree-supported schools. Schools were eligible for the study if they they met specific Mango Tree program criteria including: having two P1 classrooms and teachers, having desks and lockable cabinets for each P1 class, a student-to-teacher ratio of no more than 135 during the 2012 school year in grades P1 to P3, being located less than 20 km from the CC headquarters, being accessible by road year round, having a head teacher regarded as "engaged" by the coordinating centre tutor (CCT), and not having previously received Mango Tree-support. These criteria were deemed important by Mango Tree to support the specific aspects of the NULP instructional model. In addition, head teachers agreed to assign the two best early primary teachers in the school to the P1 classrooms. To determine eligibility, school-level data were collected from each school in late 2012. Out of 99 total

[4] If this did not yield at least 50 pupils, research assistants proceeded through the list of all remaining pupils and selected every seventh one.

schools, 38 met these criteria. Each head teacher signed a contract with Mango Tree outlining the guidelines for participation in the evaluation. These contracts had credibility: Mango Tree had used them in previous years in schools where it was piloting the NULP, and schools that did not adhere to the contracts lost Mango Tree support. All schools adhered to the contracts in 2013, so the contracts did not lead any of them to be removed from the study.

Selection of Students

During the first two weeks of the 2013 academic year, enumerators collected enrollment rosters from the P1 classrooms of each school in the study. From these rosters, we generated an ordered list of 70 randomly-selected students, stratified by classroom and gender. Baseline exams were conducted during the third and fourth weeks of school (described below). The first 50 students on the list from each school who were present in the school on the day of the baseline exams were selected into the sample.[4] These 1900 students from the 38 study schools comprise our *baseline sample.*

3.2 Randomization

The 38 schools in the study were assigned to one of three study arms via public lottery: control schools, Mango Tree-administered program schools, and Government-administered program schools. Prior to the lottery, the schools were grouped into stratification cells by the researchers based on the schools' CC, total P1 enrollment, and distance to the CC. The lottery – held publicly at a stakeholder meeting – proceeded separately for schools in each stratification cell with representatives drawing tokens indicating treatment status from an urn. We discuss tests for balance of baseline sample characteristics across treatment arms below.

4 Data

Our primary learning outcomes are measured by a set of examinations conducted at the beginning and end of the school year to assess student performance in reading and writing Leblango, and in speaking English. These data – as well as surveys among students and their parents – were collected among our baseline sample of 1900 students. In addition, we use data from teachers surveys, and classroom visits that collected attendance, enrollment, and conducted classroom observations. The remainder of this section first describes the data sources and then presents summary statistics from the baseline exams.

4.1 Student Examinations

Baseline tests were conducted in the third and fourth week of the school year among the baseline sample of 1900 students. Endline tests were conducted during the last two weeks of the school year, in

[8] The beginning instructions for the test are explained in Lango, and the tests themselves are conducted in English, with the examiner asking, for example, "What can you see?" (for subtest 3). As with the EGRA, the oral English examinations were conducted one-on-one with the students by trained examiners (they immediately followed the EGRA for each student).

late November 2013. Of the students tested at the baseline, 78 percent were also found for endline exams. This gives us a *longitudinal sample* of 1481 students, which we use in our main student analysis (attrition across treatment arms is discussed below).

Exams were administered by trained examiners hired specifically for the testing process. Examiners were not otherwise affiliated with Mango Tree, and were blinded to the study arm assignments of the schools they visited. Two of the tests, the EGRA and the Oral English Test, were conducted one-on-one by examiners sitting with individual students, making use of visual aids. The examiners marked each question correct or incorrect during the exam. The third test, the Writing Test, was conducted in a group setting with a single examiner handing out materials and instructing pupils to write a story. We describe each of the tests in detail below.

Early Grade Reading Assessment (EGRA)

Our main outcomes of interest come from the Early Grade Reading Assessment (EGRA). The EGRA is an internationally recognized exam designed to serve as an "assessment of the first steps students take in learning to read: recognizing letters of the alphabet, reading simple words, and understanding sentences and paragraphs" (RTI International, 2009). It has been adapted to dozens of languages and implemented in nearly 70 countries around the world (USAID 2014). In 2009, it was adapted to Luganda and Lango and used in Uganda to assess the reading ability of 2000 students in 50 schools across the country. We use this same adaptation of the EGRA to Lango, which covers six components of reading ability: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The first four components involve students attempting to read letters, sounds, and both real and invented words from tables that are shown to them. The last two have students attempt to read a simple passage aloud and then answer comprehension questions about it. Because Mango Tree's main teaching objective in P1 is for students to learn the names of the letters of the alphabet, the letter name knowledge component of the test is of particular interest in evaluating the success of the program.

Oral English

The eventual goal of both the standard government curriculum and the NULP is for students to successfully transition to English by P5. One potential question about local language-first education is the extent to which it increases or inhibits students' progress in learning to speak, and eventually to read and write, in English. We therefore administered a simple oral examination – designed by Mango Tree – that asks students to answer basic English vocabulary questions based on pictures. The oral English examination has three sections. The first focuses on vocabulary and counting skills, asking students to point to a specific object in a picture named in English, and count how many there are. The second section evaluates students on their vocabulary and sentence structure abilities, asking them what a specific person in a picture is doing and what the name of a particular object is. The third section is more open-ended – it presents students with a picture of a scene and asks them what objects and which people they can see in the picture.[8]

In addition to measuring students' ability to speak English, we also wanted to capture the effects of the program on students' ability to read English words. The endline exams therefore added an additional test which asked students to read a list of eighteen words commonly taught in P1 (in the standard government curriculum). Rote memorization of how to read basic words in English aloud is a

common technique in P1 classrooms in the Lango sub-Region. The NULP contrasts sharply with that practice, and does not teach any English reading during P1.

Writing

To capture improvements in students' ability to write, we made use of a writing test designed by Mango Tree and previously used to monitor writing skill acquisition in their pilot-testing of the NULP. Students completed the tests at the schools and were scored off-site by an expert in writing acquisition among children in the Lango sub-Region. The test has two broad sections. In the first section, students are asked to write their names.[9] Langi names are divided into an African surname, typically written first, and an English given name, typically written second. Surnames come from a small set of names that are passed down within extended families, with a known spelling in the Leblango orthography. Given names also come from a small list of names with known spellings. Each name was score separately in two categories: spelling and capitalization. Ability to write one's name is a major goal that Mango Tree sets for P1 students in terms of writing acquisition.

In the second section of the test, students were asked to write a story about what they like to do with their friends, and to draw a picture to illustrate the story. The picture was unscored, but served to keep children occupied who could not write anything. The story was scored in seven categories: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation.[11]

Combined Exam Score Indices

Our main learning outcomes are measured by the endline exams: reading, using the EGRA, English speaking, using the Oral English Test, and Leblango writing, using the Writing Test. Each of these exams has several modules, designed to test distinct but aspects of a child's ability rather than to produce a single overall score. The modules differ in their number of questions and some are scored based on a student's speed while others are untimed. We present the effects on each module separately, but a key question is whether the program has overall effects on each test – and how large those effects are. One challenge is that while there are guidelines for scoring each section of the EGRA, there is no defined system for combining the scores. The same issue holds for the other two tests. To measure the effect of the program on students' overall exam performance, we construct a principal components score index by normalizing each of the test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006). Our results are robust to alternative methods of index construction.[12]

[9] This is a purely evaluative exercise; exams were matched to students using pre-printed ID numbers.

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

[12] Our estimated effects for the EGRA and the Writing Test are still statistically significant, and slightly larger, for an alternative index that takes the unweighted mean across test modules, following Kling, Liebman, and Katz (2004). The estimated effect on the Oral English Test is nearly unchanged. To make these alternative indices, we normalize each module's endline score against the control-group endline score distribution for that module. We then take the simple average of the normalize scores across all the modules.

4.2 Surveys

Our analysis also makes use of two surveys, one for students and the other for teachers. Both surveys were conducted at the same time as the endline student examinations. The student surveys were a brief set of age-appropriate questions that asked them about their attitudes toward school, their effort, and their perceptions of their own ability and performance. Teacher surveys were designed to capture basic demographic details, as well as attitudes towards school and local language education. The teacher surveys also included details about teaching history, duties at the school, and time use.

4.3 Classroom Visits

Attendance and Enrollment

In addition to the baseline and endline examinations at each school, enumerators were also sent to each school three times during the school year to collect additional supporting data on the intervention. These visits took place in July, August, and October, so two visits occurred during the second term of the school year, and one occurred during the third (and last) term of the year. During these visits, enumerators collected data on attendance for all students in P1, as well as data on any new student enrollment. Attendance data was collected using the enrollment rosters. Enumerators noted whether each student on the list was present.

Classroom Observations

During the same visits at which they collected the attendance and enrollment data, enumerators also conducted classroom observations. These were detailed observations of two lessons in each of the school's two classrooms. These observations captured information about teaching strategies, student behavior and engagement, discipline, language of instruction, and a breakdown of the focus of each lesson on different topics. Enumerators were sent to the schools with paper forms with check boxes to note basic details about the school and classroom, as well as detailed information on each 30-minute lesson. School and classroom details included the teacher's name, number of students in the class, teaching and learning materials that were in the classroom, and which lesson was observed.

The details about the lesson were broken up into three 10-minute blocks. For each block, the enumerator captured the start and end time, and ticked boxes to indicate that a teacher had engaged in a range of actions during the block such as referring to the teaching guide and ignoring off-task students. They also noted the share of time the teacher spent speaking English and Leblango.

In addition to capturing details about teacher behavior, the enumerators also recorded student actions in three categories: reading, writing, and speaking/listening. Enumerators indicated the number of minutes (out of the 10 in the block) spent on each category and the share of students participating in the activity. They then ticked boxes to note whether they saw students do various actions, such as doing the activity in a group or on their own, using a specific material such as a slate for writing or a reader for reading, and whether English or Leblango was used.

4.4 Baseline Characteristics

Tables 2 and 3 presents baseline summary statistics. We focus on the first column of Table 2, which presents the mean of each variable among the control group, and column 2 of Table 3, which shows the share of students who got any answers right on each component of the EGRA. The sample is slightly less than half male and the mean age at the beginning of P1 is 7. Very few students got any correct answers on the baseline EGRA – just 40% got a single question right on the entire exam. Looking to the individual components, only 15% could identify a single letter of the alphabet, and even lower proportions scored any points on the more-advanced reading skills.[15] One notable exception to this pattern is the Reading Comprehension questions, which have the highest proportion of students getting a question right at 30%.[16] Students were even less successful on the Writing Test: more than three quarters scored zero points on the entire exam. Scores were higher on the Oral English Test, probably because it involved no reading and thus relied on skills that students might have already begun to develop before beginning school.

5 Empirical Strategy

5.1 Main Econometric Approach

Our main outcomes of interest are student performance on three exams: the EGRA, the Oral English Test, and the Writing Test. For each exam, we examine effects on each component separately, as well .

Our empirical strategy relies on the randomized assignment of schools to the three study arms for identification: randomization guarantees that the students in the three study arms will be balanced, in expectation, on observed and unobserved pre-treatment variables, allowing us to attribute any post-treatment differences in outcomes to the effect of the program the school received. While the treatment was assigned at the school level, our main analyses focus on student-level outcomes. We run regressions of the form:

$$(1) \qquad y_{is}=\beta_0+\beta_1 MTSchool_s+\beta_2 GovtSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\eta\, y^{baseline}_{is}+\varepsilon_{is}$$

Here i indexes students and s indexes schools. y_{is} is a student's outcome at endline – typically his or her score on a particular exam or exam component. $\boldsymbol{L}_s$ is a vector of indicator variables for the stratification

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

[16] This is higher than the share who were able to correctly read any of the words from the passage aloud. This may be because students are better able to make words out on the page than to correctly pronounce them out loud, and also may the result of lenient scoring by the examiners. This pattern is identical across study arms.

[18] See INSERT WEBSITE WHEN PUBLIC for details.

group that a school was in for the public lottery that assigned schools to study arms; we control for them, following Bruhn and McKenzie (2009), to increase the precision of our estimates. *MTSchool* and *GovtSchool* are indicators for the school being in the Mango Tree- or Government-administered version of the program, with the omitted category being in the control group. ε_{is} is a mean-zero error term. To account for the fact that the treatment was randomized at the school level rather than at the student or teacher level, we uniformly report standard errors that are clustered by school.

β_1 and β_2 are our estimates of the effects of the MT and CCT programs, respectively, on exam scores. To restate the identification assumption above in terms of the variables in our estimating equations, The key assumption necessary for our estimating equations to yield consistent estimates of β_1 and β_2 is that *MTSchool* and *GovtSchool* are independent of the error term ε once we condition on the other controls in the regression. This is guaranteed by process that assigned schools to study arms, which was random conditional on stratification cell. We next discuss baseline balance in further detail.

Our preferred specifications also control for the baseline value of the outcome variable, $y^{baseline}_{is}$, whenever possible. We do this for two principal reasons. First, we stated that this would be our preferred specification in our pre-specified analysis plan.[18] Second, it helps address the potential baseline imbalance on some of the test score outcomes described in Section 4.1 above. In practice, baseline values for the outcome variables are available only for the student test scores. Therefore, we include this control only in our test score regressions. We also show that our results are not materially affected by the exclusion of this control.

In addition to using equation (1) to estimate the effects of the two NULP variants on test scores, we also use the same specification to study its effects on student aspirations,

5.2 Baseline Balance

Table 2 provides evidence of balance across the study arms. The three sets of columns present means by study arm for three different samples of students: the baseline sample, the longitudinal sample, and the set of students who were lost to followup. We formally test for differences between study arms by estimating

$$(2) \qquad y_{is}=\beta_0+\beta_1 MTSchool_s+\beta_2 GovtSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\boldsymbol{\tau}\, T_s+\varepsilon_{is}$$

Here we control for $\boldsymbol{L}_s$ for the same reasons noted above. We also control for the date of the baseline exams, T_s, because it is not balanced across study arms, and because there is evidence of a time trend in scores on Oral English Test and the Writing Test, possibly because the examiners gained experience administering the tests. Statistically significant differences are indicated by stars next to the Mango Tree Program and Government Program means.

A comparison of the first three columns shows that the baseline sample is relatively well-balanced across study arms. There are no significant differences in demographics: the sample is slightly less than half male and seven years old on average at the beginning of P1. The PCA indices for the exam scores show that overall test performance is roughly the same across study arms. Looking at the detailed list of test components, however, there is evidence of a small degree of imbalance. The Government Program performs slightly worse than the control group on the Reading Comprehension ($p<0.05$) of the EGRA, while both versions of the program score somewhat lower than the control group on two of the Oral

English Test components. Students in the Mango Tree program score significantly better on the portion of the Writing Test that asks them to write their African names.

Columns 4 through 6 replicate columns 1 through 3, but for the longitudinal sample that we actually use to analyze the NULP's effects. Comparing the coefficients and statistically-significant p-values, we see that the same patterns hold for this sample as for the baseline sample: it is balanced on demographics and overall test performance, but with some significant differences in the individual test components. Columns 7 through 9 present variable means by study arm for the set of students who were lost to followup – members of the baseline sample who are not in the longitudinal sample. This sample uniformly performs worse on the baseline tests than the longitudinal sample does. This pattern is balanced across study arms in terms of the overall test score indices, but there is some evidence of differences in performance among attritters on certain test components. However, these differences are not large enough to lead to change the pattern of imbalance for the longitudinal sample relative to the baseline sample.

The small degree of imbalance in baseline test scores could have arisen from three sources. First, the random assignment of schools to study arms, which generates balance on all observed and unobserved variables in expectation, could led to an imbalanced sample in realization. Second, the same applies to the random samples of students within schools. Militating against these possibilities somewhat is the fact that the sample looks balanced on demographic factors. A third possible source of imbalance is that the baseline exams took place after the school year had begun, and so they may have picked up some initial, short-run effects of the treatment. The direction of the differences across study arms is consistent with what we would expect from the NULP's emphasis on the use of Leblango instead of English and is focus on teaching students beginning writing skills. The small amount of baseline imbalance in our sample motivates our choice to control for baseline values of the outcome variable in all our test-score regressions.

5.3 Additional Specifications

We supplement the student-level analyses in equation (1) above with several others. First, we use the set of classroom observations. In these, each school in the study was visited three times; during each visit, both classrooms in the school were observed during two separate lessons. To analyze these data we estimate:

$$(2) \qquad y_{lrcs}=\beta_0+\beta_1 MTSchool_s+\beta_2 GovtSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\mathbf{R}'_r\boldsymbol{\delta}+\mathbf{E}'_{rcs}\boldsymbol{\rho}+\mathbf{D}'_{rcs}\boldsymbol{\mu}+\varepsilon_{lrcs}$$

Here *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) above, equation 2 adds as controls vectors of indicator variables for the round of the observation ($\boldsymbol{R}_r$), the enumerator conducting the observation ($\boldsymbol{E}_{rcs}$), and the day of week of the observation ($\boldsymbol{D}_{rcs}$).[19] ε_{lrcs} is a mean-zero error term.

Enrollment data is collected as total numbers at the school level, so we analyze it at the school level as well:

$$(3) \qquad y_s=\beta_0+\beta_1 MTSchool_s+\beta_2 CCTSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\varepsilon_s$$

[19] The classroom observation results are nearly identical in magnitude but less precise in magnitude when we omit the enumerator and day-of-week fixed effects.

Here *s* indexes schools, ε_s is a mean-zero school-level error term, and all other variables are defined in the same way as in equation (1). We also examine the sensitivity of our results to using the log of enrollment instead of its level.

We use information from the endline teacher surveys to study how the program affected teacher's effort (time use, interactions with parents) beliefs and attitudes, and participation in training. To study these, we estimate program effects at the teacher level by estimating:

$$(4) \qquad y_{js}=\beta_0+\beta_1 MTSchool_s+\beta_2 CCTSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\varepsilon_{js}$$

where *j* indexes teachers and *s* indexes schools, and ε_{js} is a mean-zero teacher-level error term; all other variables are defined as in equation (1).

6 Results

Our analysis first focuses on the effects of the two program variants on student exam scores. First, as a benchmark, we discuss the performance of P1 students under the status quo government curriculum – that is, student performance at the endline in control schools. We then turn to impacts on the EGRA, the Oral English Test, and the Writing Test.

6.1 *Status Quo* Performance in Literacy at the end of P1

In addition to its use in measuring the impact of the NULP on literacy, the exam data we collected allows us illustrate the gains P1 students in the Lango sub-region make in terms of reading ability in the absence of the program. The blue bars in Panel B of Figures 2 and 3 show how students in the control schools performed on the EGRA at the end of P1; these changes are also summarized numerically in columns 5 and 6 of Table 3. At the end of one year of school, roughly 50% of students could not recognize a single letter of the alphabet (Figure 2 Panel B). Just over 20% could recognize between one and five letter names, and a similar fraction could recognize between six and twenty. Fewer than 10% of pupils could correctly identify more than twenty letters out of a total of 100 chances.

The NULP sets learning the names of letters as a key goal for P1 students, arguing that it is a critical building block for more-advanced reading skills. Consistent with this claim, overall reading performance mirrors the performance on letter-name recognition. The blue bars in Panel B of Figure 3 show that 40% of all students could not answer a single question correctly on the entire EGRA. The remainder of Figure 3 Panel B confirms that overall EGRA performance is largely driven by letter name recognition in P1.

A comparison between the first and second panels of Figures 2 and 3, focusing on the blue bars, reveals a staggering lack of improvement in reading over the course of P1. Over 80% of students enter P1 unable to recognize a single letter of the alphabet, and the majority of those students leave P1 having made no progress whatsoever. Overall EGRA scores do not look much better: 40% of students get at least one correct answer across the six components of the exam at the beginning of the school year, but that number rises to just 60% by the end of the year. A small number of highly-performing readers do much better than the typical student: the fraction of students, answering more than twenty questions right rises from negligible at the beginning of the year to 10% by the end of the year. But these top students leave the preponderance of their classmates far behind.

The measured increases in exam scores in the control group form a natural basis for comparison for the effects of the two variants of the NULP on exam scores: we can compare the gains from the program to the typical gains experienced by a child during P1. We now turn to the impacts of the program on the EGRA, performance on which is our main outcome of interest.

6.2 Program Effects on EGRA Scores

The impacts of the two versions of the NULP on EGRA scores are shown in Table 4, which estimates equation (1). Column 2 presents the impact on students' knowledge of letter names, the principal learning goal that Mango Tree sets for P1 students. The Mango Tree-administered version of the program has a very large impact on letter name knowledge: scores increase by 1.01 standard deviations. The government-administered program improves performances in recognizing the names of letters by 0.41 SDs, which is still a significant gain but less than half as much as the full-cost version of the program.

Examining the effects of the two versions of the program on the other EGRA components reveals a more nuanced picture. The Mango Tree-administered program has strong effects on all six components that are uniformly significant at the 0.05 level. The government-administered program, however, has no statistically-significant effect on any EGRA component other than letter name knowledge. The low-cost version of the program, then, improved only the headline measure of literacy emphasized by Mango Tree, with no benefits to other, more advanced aspects of literacy.

This finding is verified by Column 1 of Table 4, which presents estimates for the combined score index described in Section 4.1 above. The Mango Tree-administered program raises this index by 0.63 SDs, confirming that the large effect of the program on exam scores is not merely an artifact of focusing on knowledge of letter names. Even taking 0.63 SDs as our best estimate of the program's impact on reading ability, the effect of this program would be among the largest ever measured in a randomized trial of an education program (McEwan, 2013). Moreover, we can reject gains smaller than 0.37 SDs at the 0.05 level; in the few cases where large effect sizes have been found in primary education programs, those effects have had wide confidence intervals that do not exclude much smaller impacts. The government-administered program's effect on the EGRA index is just 0.13 SDs and is statistically indistinguishable from zero. The estimated effects on EGRA performance are virtually unchanged when we omit the baseline exam score controls; see Appendix B.1 for a detailed discussion and tables. The huge magnitude of the benefits of the program for reading is evident from Panel B of Figure 2. It shows the distribution of endline letter name knowledge scores by study arm. The full-cost version of the NULP cuts the share of students that cannot recognize a single letter in nearly half, and nearly quadruples the share that can recognize 21 or more letters. The effects are similarly clear-cut in Panel B of Figure 3, which shows the distributions of the total number of points scored on the EGRA. The low-cost variant of the NULP achieves smaller improvements in both letter name recognition and overall EGRA performance. It shifts the score distribution to the right, but does so by a smaller degree than the full-cost variant.

6.3 Program Effects on English Speaking and Word-Recognition Ability

Since the NULP focuses on promoting the use of the local language, Lango, in classrooms, one area where the program could potentially have effects is on students' English speaking skills. One concern parents and other stakeholders in the Lango sub-Region have expressed with mother-tongue curriculum is that it would crowd out English skills. Table 5 presents the effects of the two program

variants on students' scores on the oral English examination, estimated using equation (1). Neither the Mango Tree-administered nor the government-administered version of the program had a robustly statistically-significant effect across the different examination components. Column 1 shows that the overall effect of the NULP on the combined score index is statistically insignificant for both program variants. The Mango Tree-administered version raises this index by 0.14 SDs, and the Government-administered version lowers it by 0.09 SDs.

Although the overall effect of the program on English speaking ability is not statistically significant, the point estimates in the table still represent our best estimate of the effect of the program; these are uniformly negative for the government-administered program but mostly positive for the Mango Tree-administered version. Moreover, Columns 8 and 9 show that the Mango Tree-administered program had statistically-significant benefits for the third subtest, expressive vocabulary, which uses relatively open-ended questions about a scene ("What do you see?" and "Who do you see?") as opposed to the naming of specific objects and actions ("What is this?" "What is she doing?"). This is noteworthy because the *status quo* in P1 classrooms in the Langa sub-Region is to focus on the rote memorization of English words, as opposed to actual usage; while control-school students might have an automatic advantage on the closed-ended questions, NULP students are more likely to have gained on open-ended questions. The estimated effect of the Mango Tree-administered version of the program on students' expressive vocabulary is roughly 0.3 SDs for each of the two subtests, which provides suggestive evidence that, in addition to reading Langa, the program also improved students' actual English speaking ability.

This argument is also buttressed by Column 10, in which the outcome is a separate test in which students were asked to read a set of 18 printed English words aloud. This is a task that the NULP does not have teachers spend any time on in P1, because English reading does not commence until P2. However, it is common in *status quo* classrooms in the Langa sub-Region. The test was designed to use words that are commonly used in English curricula in P1 classes; it thus captures the extent to which students have either actually learned to read these words in English or have memorized by rote what to say when they are pointed to. NULP students perform substantially worse on this task, by 0.21 SDs under the Government-administered version and by 0.29 SDs under the Mango Tree-administered version. The latter estimate is significant at the 0.05 level. This result, along with the results from the Oral English Test, suggest that there is no evidence that the NULP harms students' progress in learning English. While they do worse on a simple rote memorization task, they actually improve substantially in their ability to use English in an expressive and open-ended manner.

6.4 Program Effects on Writing

We examine the effect of the two versions of the program on writing ability in Table 6, which shows impacts on Mango Tree's writing test, estimated using equation (1). Columns 2 and 3 show that both versions of the program have large effects on the first section of the exam, which asks students two write their first and last names. Learning to write one's name is the main goal of the NULP for P1 students. The Mango Tree-administered program also has positive effects on the second section, in which students are asked to write a short story (Columns 4 to 10). The combined writing test index rises by 0.42 SDs (Column 1), which is statistically significant at the 0.05 level. The government-administered program, however, has uniformly negative effects on the story-writing component of the exam, with the negative effects on Voice, Word Choice, and Presentation reaching significance at the

p=0.05 level.[20] The combined Writing Test score index falls by 0.17 SDs, although this drop is not statistically significant. This suggests that the government-administered version of the program significantly boosted the headline measure of writing ability – name writing – at the cost of progress in overall writing skills, and in particular the ability to actually write a passage.

7 Mechanisms of the NULP's Effects

Tables 4 through 6 illustrate that the full-cost version of the Mango Tree program has significant benefits for pupil literacy, with some evidence of ancillary benefits for English-speaking ability, while the reduced-cost version seems to achieve gains on only the most basic outcomes that are targeted as goals for P1 students – letter recognition and name writing, with no gains in other areas and statistically-significant losses on more advanced aspects of writing ability. The two variants of the program were randomly allocated as complete packages, so we cannot causally separate which parts of the program had the most benefits or where the downsides of the low-cost version are coming from. However, we can approach the question of why the program worked, and why the lower-cost version backfired in some areas, by looking at evidence on intermediate outcomes that may shed light on the program's mechanisms.

In this section we discuss each set of intermediate outcomes in turn: the student surveys, the classroom observations, attendance and enrollment, and teacher surveys. We then draw general conclusions about what all these data sources tell us about the mechanisms behind the NULP's impacts on learning.

7.1 Changes in Student Effort, Beliefs, and Attitudes

To do this we begin by looking at students' responses on the age-appropriate surveys that we conducted during the endline exams. The effects of the two program variants are shown in Table 7. The effects are estimated using equation (1), but without controlling for baseline values of the outcome because no data was collected on these outcomes at baseline. Students in both versions of the program show evidence of increases in perceived ability. They are more likely to report that they think they will pass the PLE (primary leaving examination), a high-stakes test that determines secondary school admissions, at the end of primary school. The estimated increase is 2.2 percentage points for the Mango Tree-administered program and 1.5 percentage points for the government-administered program (column 1), over a very high base rate of 95%.[21] Likewise, students' perceived class rank improves by 0.15 SDs in the Mango Tree-administered program (no effect is seen for the government-administered version).

We find mixed results on enthusiasm for school and future aspirations. No effects are evident on students preferring school to other activities or preferring literacy class to math (columns 2 and 3); the estimated effects are not just statistically insignificant but nearly zero in magnitude. However, we do see evidence of admiration for teachers and an appreciation for education: students in the Mango Tree-administered program are seven percentage points more likely to want to go into a career in education

[20] One of the 12 control schools was mistakenly instructed to complete the Writing Test in English instead of Leblango. Our results include this school, with the test marked in English. Our findings are robust to dropping the stratification cell for this school from our sample – see Appendix B.2 for a detailed discussion.

[21] The actual pass rate is much lower: in most Ugandan schools, fewer than half of students who begin P1 even complete P7 and take the PLE, and a small fraction of those pass it.

(column 4). This is offset by an eight percentage-point drop in desire to become a doctor or nurse (column 5). Since students could list only one career, and the NULP does not affect how amibitious of a career students want (column 9), this suggests that the most ambitious students in class now want to go into education instead of healthcare.

Finally, a roughly zero effect is also seen for our measure of effort, practicing writing at home (column 6) This suggests that changes in student effort in literacy are not important drivers of the observed effects. Overall, the results from the survey suggest that there was some increase in student confidence and enthusiasm for school, and these effects are larger for the Mango Tree-administered program than for the government-administered version. This gap may help explain part of the gap between the impacts of the full-cost and reduced-cost versions of the program on student test scores.

7.2 Changes in Teacher and Student Behavior in the Classroom

The most likely mechanism for the program's effects is that it changes how teaching actually takes place in the classroom. To explore this, we examine data from a set of classroom observations that measured teacher (Table 8) and pupil behaviors (Tables 9, 10, and 11) during class. These four tables use regressions of the form specified in equation (2). Table 8 reveals that both variants of the program induced teachers to spend more of their time speaking in Lango, by twelve percentage points for the full-cost NULP and nine percentage points for the reduced-cost variant (Column 2). Teachers in the full version of the program were also more likely to move around the classroom – they were twelve percentage points less likely to simply remain at the front of the class (significant at the p=0.05 level) and nine percentage points more likely to move freely throughout the classroom (not statistically significant). Teachers in both NULP variants were 6 percentage points more likely to be observed ignoring off-task students (Column 7), with no statistically-significant changes in the other outcomes. This is somewhat surprising, but it may reflect the establishment of a better overall classroom environment: in an ideal classroom full of readily-participating, on-task students, teachers will never have to bring students back onto task. Also, teacher training courses often encourage teachers to ignore off-task students rather than call attention to them.

Table 9 shows differences across study arms in student behavior while working on reading tasks. Students in both versions of the NULP are more likely to be observed reading sounds, and students in the full-cost version are more likely to be seen reading full sentences. Both variants of the program are more likely to be reading out of readers or primers. The proportion of reading done in Leblango rises by 22%. Classes also spend a higher proportion of time on reading: an additional 0.7 minutes per ten-minute observation window for the full-cost version of the program, and 0.5 minutes for the reduced-cost version. This represents an increase of roughly 15% over the control-group mean of 3.7 minutes.

In Table 10, we examine the changes in student behavior while writing. Students in the full-cost version of the NULP are 8 percentage points more likely to be observed drawing pictures, and 6% more likely to spend time writing their names. The Government-administered program shows a 6 percentage-point increase in the chance students will be seen air-writing, but there is no comparable effect for the Mango Tree-administered program. This may reflect the fact that the Government-administered version of the program did not include the writing slates. If students lacked their own exercise books to write in, this would force teachers to improvise if they want to their students to be able to practice writing. Another difference that is asymmetric across program variants is a 9% rise in the chance that students in the Mango Tree-administered version will be seen writing their own text. This is a gain of more than 100% over the control group mean, and helps explain the large

improvements in passage writing in that version of the program. The change in the amount of time spent on writing is not statistically significant, but is comparable in magnitude to the increase in time spent on reading: about 16% of the control-group mean of 1.2 minutes for the Mango Tree-administered version of the program, and 29% for the Government-administered version.

Finally, Table 11 turns to changes in student behavior while speaking and listening. Students more than double the chance that they speak or listen in small groups, and the chance that students will be observed speaking and listening to the teacher falls by a comparable magnitude. This is consistent with a drop in the amount of rote memorization "call and response"-style learning that is typical in *status quo* schools in the Lango sub-Region. The share of speaking and listening done in Lango rises, which would fit into a story where students especially spend less time doing rote call-and-response in English, to memorize English words. Finally, the amount of time spent on speaking and listening falls by 16% of the control-group mean in the full-cost version of the program (significant at p=0.05) and by 7% in the reduced-cost version (not statistically significant). This also matches a story in which the teacher engages in less call-and-response repetition of words and phrases as a way to memorize them.

7.3 Changes in Attendance and Enrollment

Teacher and student behavior during class can be thought of as variation at the intensive margin of effort. Another important factor is changes at the extensive margin: whether students and teachers show up for class at all. Table 12 shows estimated differences in student attendance and enrollment and teacher attendance across study arms. Columns 1 to 4 are estimated using equation (1) on the full sample of students enrolled in the schools at baseline. Column 5 is estimated at the school level using equation (3). Column 6 is estimated at the teacher level, using equation (4). There is no evidence of any differential changes in enrollment across study arms, nor of differences in teacher attendance. There is some evidence of a limited increase in attendance for students in the Mango Tree-administered version of the program (it rises by 5 percentage points, with $p<0.1$), concentrated in the first visit to schools which happened early in the second term of the school year. Students in the Government-administered version of the program are 4 percentage points less likely to attend than control-group students.

Though the p-value on this difference exceeds 0.1, the difference from the full cost version of the program is statistically significant, and is 9 percentage points over a base of 42% attendance. The lower attendance is concentrated toward the end of the school year. Part of the improvement in performance in the Mango Tree-administered version the NULP may be due to the simple fact that students are exposed to more teaching because they were in class for longer. The smaller gains in the low-cost variant of the program can be ascribed in part to students spending less time in class than in the full-cost variant.

7.4 Changes in Teacher Effort, Beliefs, Attitudes, and Training

Our final ancillary data source for examining the mechanisms of the NULP's benefits is the endline teacher surveys, which were done at the same time as the endline exams. We estimate effects on the survey outcomes using equation (4). The outcomes are grouped into three categories: columns 1 to 5 measure teacher effort; columns 6 to 10 measure teacher beliefs and attitudes, and columns 11 to 14 measure the main human capital input the NULP provides for teachers – training.

Changes on teacher effort as a result of the program are fairly muted. The full-cost version of the program shows a marginally-significant increase in the amount of time spent on helping students outside of the classroom, but it is large in magnitude – 2 hours more per week, nearly as much as the

control-group mean. There are no appreciable changes in interactions with parents: the number of parents the teacher met with during the school year is essentially unchanged, and this result is consistent with other outcome measures that we omit for space reasons. The one margin of effort where we detect effects is a significant increase in the chance a teacher has taught literacy classes (reading and writing), which rises from 61% in the control group to 80% in the Government-administered version of the program 92% in the Mango Tree-administered version. The NULP appears to reduce the division of labor across the two P1 teachers, which in the control group are more likely to split the literacy and non-literacy parts of class.

While effort changes very little, we observe large shifts in beliefs and attitudes. Both variants of the NULP cause teachers to be 20 percentage points more likely to say they would still want to teach if they could go back and re-pick their career. Though this is significant only for the Government-administered NULP, and only at p=0.10, pooling the two study arms for this outcome generates the same coefficient and significance at the p=0.05 level. Teachers in the Mango Tree-administered program are less likely to blame teachers for students' failure to learn, which could mean they feel less frustrated when their students struggle. Teachers in the Government-administered program rate themselves 0.3 points lower than control teachers do on a 1-3 scale of relative performance. Both versions of the program sharply reduce teachers' satisfaction with the reading performance of higher-year students in their schools, suggesting an elevation of standards. Consistent with this, and in contrast with the students' self-perceptions, there is no change in teachers' beliefs about their students' ability to eventually pass the PLE. The overall pattern is one of higher standards for students, and some increase in satisfaction with teaching as a career.

The effects of the program on training are interesting primarily because there is evidence of substitution of the NULP's training opportunities for other ones that teachers might do instead. There are increases in the rate of attending any training and the total days of training attended, which is sensible because the NULP invests heavily in training teachers. This is reflected by an approximately 50 percentage-point increase in having attended a training provided by an NGO (Mango Tree is perceived locally as an NGO despite its status as a private-sector business). But there is a compensating decline of roughly half that magnitude in attending other training. This may mean that some of the training Mango Tree provides for the NULP simply substitutes for other valuable human capital investments that teachers would be making anyway. However, trainings for public sector workers are often seen not as ways to invest in skills but as opportunities to earn extra income through the per diem payments that are provided. Thus declines in attending other training may actually reflect increased effort put toward the broader job of teaching students.

7.5 Overview of Potential mechanisms

In this section we summarize the findings from our four ancillary datasets to address two key questions about the mechanisms of the NULP's effects on student performance. First, how exactly does the program achieve such enormous gains in student performance in reading? Second, why did the low-cost version of the program backfire in terms of writing, leading to decreases. Our ancillary data sources allow us to identify two broad mechanisms help us answer the first question: changes in beliefs and attitudes and changes in how class time is spent.

The NULP causes marked changes in beliefs and attitudes: students become significantly more positively-inclined toward school, and teachers become marginally more positively-inclined toward teaching. Students believe more in their own ability, and teachers have higher standards for student

performance. These attitudinal factors could improve learning in two ways. The first way is that they could reduce the cost of effort, leading to higher effort and better performance. This could operate in our setting through changes in effort that we do not observe or do not measure well – how closely students pay attention in class, for example, or how much of official class time teachers actually spend teaching (since teachers are likely to teach for the whole period while actually being observed). The second way is by making learning easier for psychological reasons that do not involve any changes in effort.

We see no evidence of effects on student or teacher effort, which we mostly measure at the extensive margin – time spent on educational activities. Likewise, attendance is affected only marginally by the program. However, at the intensive margin of student and teacher effort – choices about how time is allocated within the fixed class periods – we observe large changes in behavior. More time is spent on reading and writing, and less on speaking/listening activities that probably reflect rote memorization through call-and response. Students spend more reading time on making out sounds, which helps develop a key basic skill on which literacy is built. Much more time across all lessons is spent speaking Leblango instead of English. Broadly, teachers spend more of their class working on actual reading skills and focusing on Leblango, and less time having their students repeat English words they can see on the board but have trouble attaching meanings to. In addition to contributing to the large gains in literacy the program causes, the effects of this channel are also evident in performance in English. Students in the full-cost version of the program do much worse at reading common English words aloud but much better at actually speaking English.

Our analysis of the four ancillary data sources also helps us address the second question. The larger gains in the full-cost version of the program can be ascribed partially to attendance. While the full-cost NULP did not change attendance significantly relative to the control schools, it did have significantly higher attendance than the reduced-cost version. This difference was particularly sharp toward the end of the year, which helps explain why more advanced reading did not improve in the reduced-cost version of the program, and also why writing might have actually gotten worse. The potential role of attendance raises the question of why attendance suffered in the schools that received the Government-administered version of the program. We cannot answer this question definitively, but we can raise a couple of possibilities. One is that students have gotten lost and stopped bothering to come to school. A second is that teachers have engaged in the practice, common in the Langa sub-Region, of chasing away the worse-performing children so they can focus on the better-performing students.

A second potential contributor to lower performance in the Government-administered version of the program is reduced inputs. In particular, students in the Mango Tree-administered NULP were given slates and the ones in the Government-administered version were not. In simple terms, this could be thought of as an input x into an education production function $L=L(x,y)$ that takes x (and also other factors, y) as inputs, and has positive and diminishing marginal returns to x. Schools that do not get slates ($x=0$) should then have lower levels of L, than schools with positive values of x, but there is no reason that removing the slates from the Government-administered NULP should lead to worse performance than in the control schools.

What could explain the worsening in performance in writing is that the NULP actually alters the production functions for various writing outcomes. The NULP provides tightly-organized lesson plans, with specific ways of teaching different skills. In the absence of the slates, Mango Tree assumed that schools would simply substitute the students' own exercise books for writing practice. What happens when those are also not available? The evidence from the classroom observations suggests that teachers substitute classroom time toward the parts of the curriculum that are more manageable: students in the

Government-administered program are more likely to practice "air writing", where they practice tracing out words and letters with their fingers. They do not see the increases in practicing writing their own text experienced by the students in the Mango Tree-administered program. The conclusion we draw is that resource-strapped teachers may have focused the time they spent on writing on the more-manageable parts of the NULP curriculum, and ended up spending less time in the aggregate on actual useful writing skills.

We conclude the results section with a discussion the cost-effectiveness of each variant of the program and the implications of our findings for the use of cost-effectiveness comparisons.

7.6 Cost-effectiveness

The large effects of the program naturally raise the question of its cost-effectiveness. While few other programs have shown such large gains, can the NULP compete on a value-per-dollar-spent basis? We examine this question in Table 14, which presents the cost per 0.2-SD gain and the SD gain per dollar spent for three different measures of the program's effectiveness. We begin with letter name knowledge, the most important outcome emphasized by Mango Tree for P1 students. The full-cost version of the program shows a gain of 0.7 SDs in this measure for each dollar spent, which trails the 0.9 SDs per dollar figure for the reduced-cost version of the program. Based on this outcome, it would cost an extra 56 cents per student to raise scores by 0.2 SDs.

A more detailed analysis tells a different story. The second and third panels of the table present the same analysis for the overall indices of reading and writing ability. Relying on overall reading ability instead of just letter-name knowledge reverse the conclusions in terms of cost-effectiveness: the Mango Tree-administered version of the program yielded over twice the gains in performance per dollar compared to the government-administered version. The writing ability index shows an even starker pattern: because the government-administered version of the program actually reduced writing performance, the cost per 0.2-SD gain from that version of the program is undefined. Instead, each dollar spent on the government-administered version of the program will decrease writing performance by 0.04 SDs. This finding raises general questions about the use of cost-effectiveness measures in comparing the effects of education programs: they may mask considerable heterogeneity in program impacts across educational domains, leading to relatively cheap gains that come at potentially large hidden costs.

8 Conclusion

The educational challenges facing the Lango sub-Region of Northern Uganda typify those present across rural Africa. Literacy rates are low, little learning is achieved in schools (despite recent successes in increasing enrollment), few students finish primary school, and the broader context is characterized by limited resources and a wide range of constraints on policymakers, educators, and parents. These challenges have helped lead to an increased call for cost-effective ways to promote learning in Africa. We evaluate one approach, developed by a Uganda-based company Mango Tree, that focuses on promoting literacy through native language-first instruction in first-grade classrooms in the Lango sub-Region.

We measure the impact of two variants of the program: a full-cost version, implemented by Mango Tree, and a reduced-cost version, which was implemented by government officials from Uganda's Ministry of Education and Sports. The full-cost version of the program causes large improvements in

students' reading and writing ability across all measures of each, and we find suggestive evidence of gains in English speaking ability as well. The reduced-cost version is less effective: it shows improvements in the headline measures of student reading and writing that are the basic benchmarks for first-grade students in Uganda. Our analysis suggests that the gains in both versions of the program may be partly attributable to increased student confidence and enthusiasm, and to increased use of the students' native language in class. The larger improvements in the full-cost version of the program may arise in part from teachers having better control of their classroom and encouraging more interactive and participatory lessons.

While the government-administered version of the program is less effective at improving literacy, it is much lower-cost and hence cheaper in terms of value-per-dollar for the headline measure of reading. However, this result hides significant variation in the impact of the low-cost version of the program on different measures of student performance. Students show no gains in more advanced aspects of reading and actually do worse than control schools on the advanced aspects of writing. The cost-effectiveness result is completely reversed when a more comprehensive measure of performance is used: it is the full-cost, Mango Tree-administered version of the program that provides more value per dollar in improving student performance. The cost-effectiveness of the Mango Tree-administered program is very high: at $2.76 per 0.2 SD gain in the benchmark component of the literacy exam for first-graders (and $4.41 per 0.2 SD gain for a comprehensive reading ability index) it is among the most cost-effective educational interventions to be measured in a randomized experiment (JPAL 2014). However, our findings indicate that these comparisons are highly sensitive to the outcome measure used, leading to not just small shifts in the exact figures but also total reversals in the sign of the measured gain per dollar (a switch from gains into losses).

Our results also suggest that attempting to reach more students with an intervention by reducing monetary and physical inputs can backfire in specific ways. The low-cost version of the program substantially increases scores on the headline measures of reading and writing ability for first-graders – the exact outcomes emphasized by Mango Tree in their internal assessments of how well the program is going. These gains come at a cost to other, less-prioritized measures: no gains in more-advanced reading skills were seen, and more-advanced aspects of writing actually got worse. One potential reason for this is that due to constrained resources, teachers in the reduced-cost version of the program may reduced the effort and inputs that would have gone toward the lower-priority aspects of reading and writing, in order to make sure they achieve the basic benchmarks. To the extent that this happened, it was without any high-stakes test to speak of: the results of the EGRA exams were not used in evaluating any of the teachers and were not even communicated back to them. Teachers' own intrinsic motivations, perhaps spurred by the program, were enough to cause unintended drawbacks from the program. Future research should explore the role of teacher effort and motivation to further document and understand this pattern; in addition, more research is needed to understand which components are critical to achieving the large across-the-board gains of the NULP, and which can be reduced or cut in order to deliver results in a truly cost-effective fashion.

Figure 1: Randomization of Schools to Study Arms

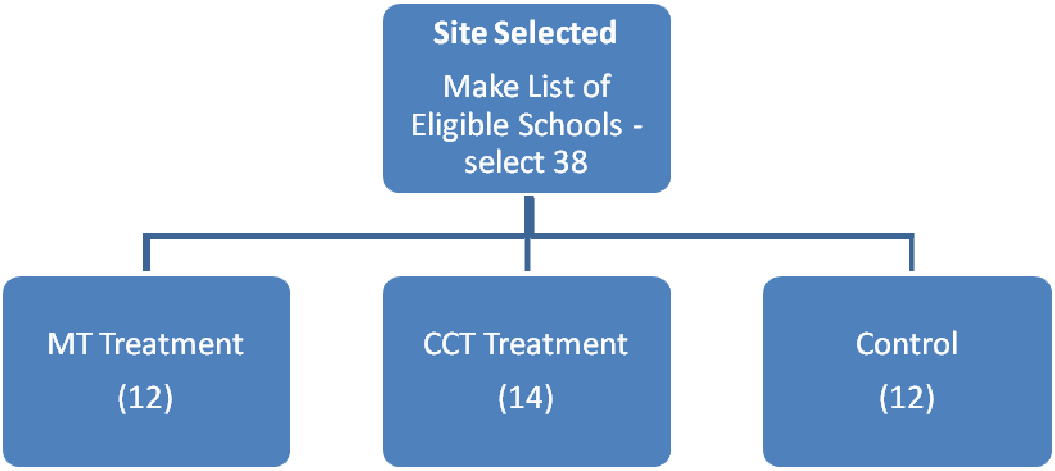
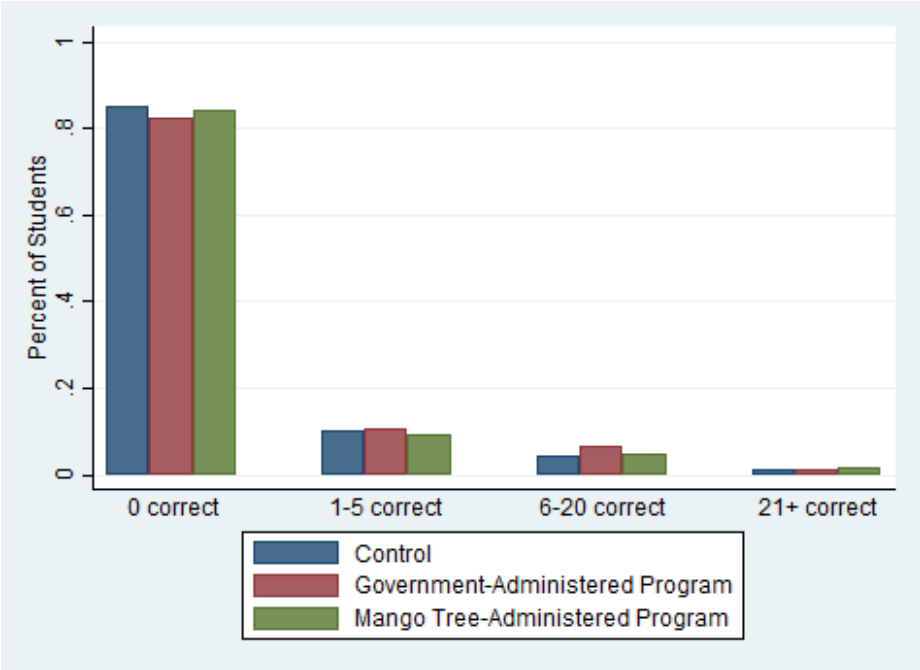


Figure 2: Performance on Letter Name Recognition by Study Arm (Number of Letters Correctly Recognized)

Panel A: Baseline



Panel B: Endline

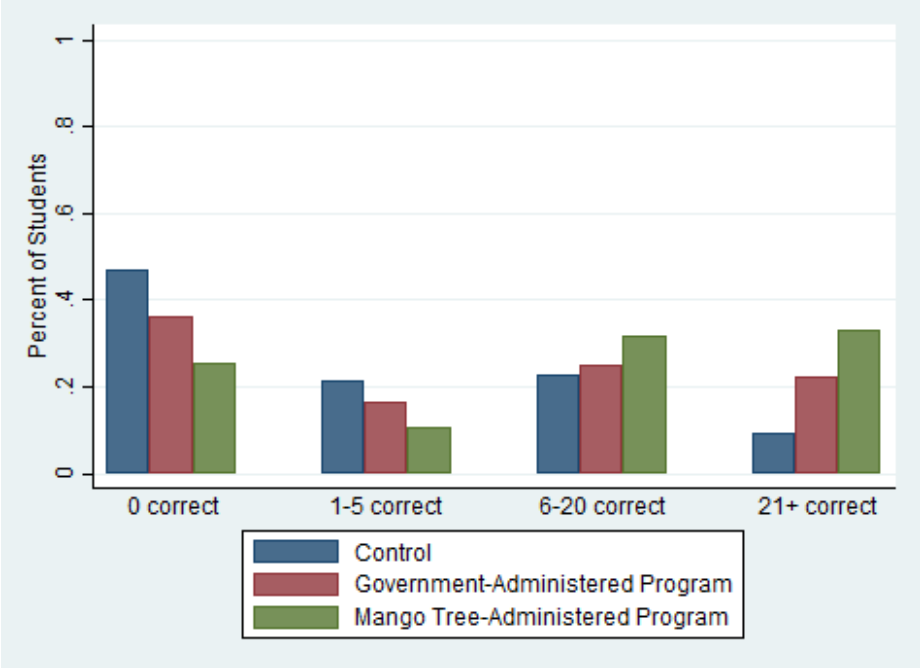
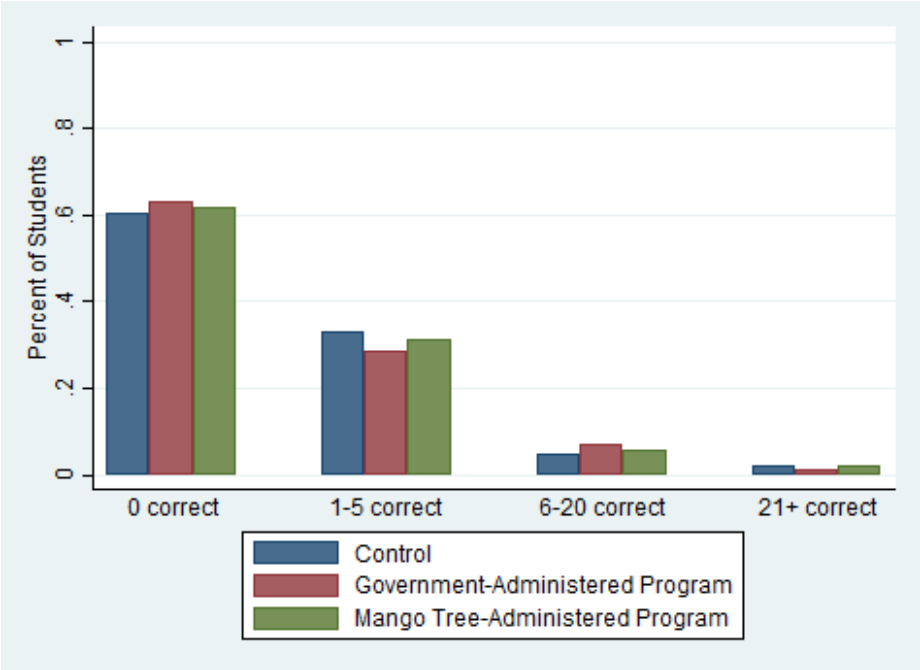


Figure 3: Performance on Overall EGRA by Study Arm
(Total Questions Answered Correctly)

Panel A: Baseline



Panel B: Endline

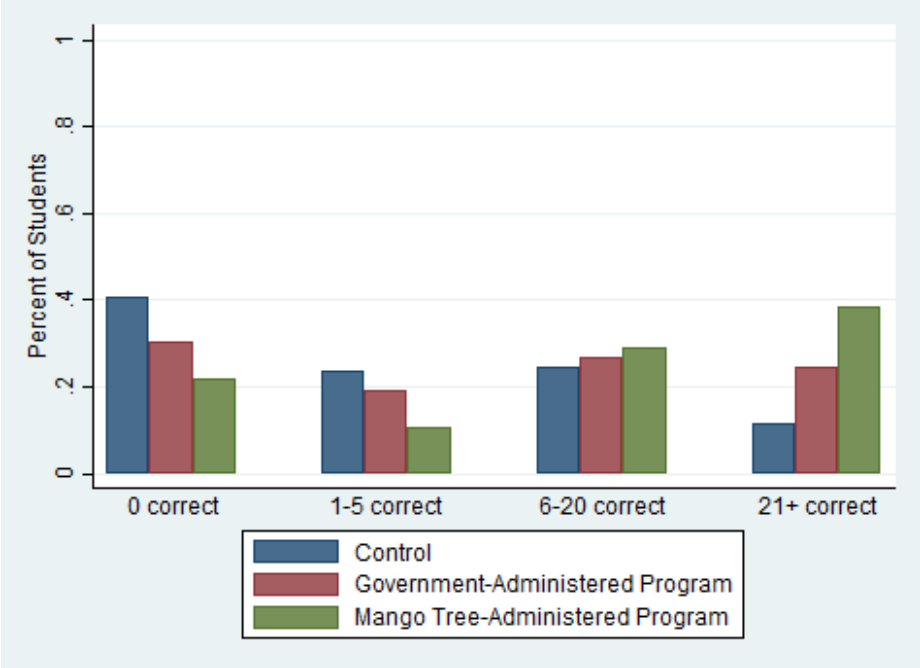


Table 1: NULP Components by Study Arm

	NULP Components Received						
Study Arm	**Slates and Wall Clocks**	**Textbooks and Primers**	**Teachers Guides**	**Training and Support**	**Parent Meetings**	**Take a Book Home Activity**	**Monthly Radio Program**
MT Program (12)	X	X	X	X (MT)	X (MT-Run)	X	X
CCT Program (14)		X	X	X (CCT)	X (CCT-Run)		X
Control (12)							X

Table 2: Baseline Covariate Balance, Longitudinal Sample

	Baseline Sample			Longitudinal Sample			Lost to Followup		
Variable	Control Mean (1)	MT Program Mean (2)	Govt. Program Mean (3)	Control Mean (4)	MT Program Mean (5)	Govt. Program Mean (6)	Control Mean (7)	MT Program Mean (8)	Govt. Program Mean (9)
Present at Endline	0.795	0.808	0.741	1.000	1.000	1.000	0.000	0.000	0.000
Male	0.486	0.509	0.474	0.488	0.524	0.479	0.475	0.447	0.460
Age	7.018	7.078	7.017	7.013	7.052	7.000	7.041	7.191	7.066
EGRA									
PCA EGRA Score Index	0.000	0.006	-0.084	0.001	0.046	-0.100	-0.003	-0.160	-0.038
Letter Name Knowledge (Letters per Minute)	1.150	1.190	1.274	1.180	1.377	1.206	1.033	0.400*	1.469
Initial Sound Identification (Sounds Identified)	0.153	0.123	0.070	0.161	0.148	0.046	0.122	0.017	0.138
Familiar Word Reading (Words per Minute)	0.169	0.182	0.044	0.168	0.225	0.025	0.171	0.000	0.099
Invented Word Reading (Words per Minute)	0.094	0.132	0.029	0.084	0.163	0.008	0.130	0.000	0.088
Oral Reading Fluency (Words per Minute)	0.503	0.552	0.126	0.508	0.684	0.037	0.480	0.000	0.382
Reading Comprehension (Questions Correct)	0.327	0.318	0.266**	0.327	0.342	0.272*	0.325	0.217	0.249
Oral English Test									
PCA Oral English Score Index	-0.000	-0.326	-0.265	0.084	-0.284	-0.244	-0.327	-0.501	-0.325
Test 1 (Vocabulary)	1.645	1.122	1.254	1.774	1.212	1.274	1.146	0.739	1.199
Test 1 (Count)	0.452	0.177**	0.276*	0.501	0.181**	0.279**	0.260	0.157	0.265
Test 2a (Vocabulary)	0.637	0.240**	0.360**	0.669	0.245**	0.391*	0.512	0.217*	0.271***
Test 2a (Phrase Structure)	0.723	0.460	0.496	0.801	0.487	0.538	0.423	0.348	0.376
Test 2b (Vocabulary)	1.328	0.797*	1.091	1.400	0.866	1.106	1.049	0.504***	1.050
Test 2b (Phrase Structure)	1.378	1.197	0.941	1.520	1.285	0.992	0.829	0.826	0.796
Test 3 (Vocabulary, Expressive - Objects)	2.188	1.657	1.763	2.365	1.724	1.802	1.504	1.374	1.652
Test 3 (Vocabulary, Expressive - People)	1.392	1.347	1.223	1.505	1.414	1.206	0.951	1.061	1.271
Writing Test									
PCA Writing Score Index	0.000	-0.024	-0.165	0.067	0.001	-0.144	-0.259	-0.130*	-0.226
African Name (Surname) Spelling & Capitalization	0.180	0.323***	0.181	0.201	0.348***	0.193	0.098	0.217**	0.149
English Name (Given name) Spelling & Capitalization	0.127	0.043	0.054*	0.145	0.043*	0.058*	0.057	0.043	0.044
Ideas	0.005	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000
Organization	0.002	0.002	0.000	0.002	0.002	0.000	0.000	0.000	0.000
Voice	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Word Choice	0.057	0.023	0.016	0.069	0.023	0.019*	0.008	0.026	0.006
Sentence Fluency	0.005	0.000*	0.001	0.006	0.000*	0.002	0.000	0.000	0.000
Conventions	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Baseline Sample includes 1,900 students who were tested at baseline. Longitudinal Sample includes 1,481 students who were tested at baseline as well as endline. Lost to Followup includes 419 students who were tested at baseline but not at endline. Stars indicate cluster-adjusted p-values for a test of the null hypothesis of no difference between each NULP variant and the control group, conditioning on stratification cell indicators and the date of the baseline exam: * p<0.05, ** p<0.01, *** p<0.001.

Table 3: Improvements in Test Performance Over the School Year, Control Group

		Baseline			Change from Baseline to Endline	
Variable	N (1)	% Any Correct (2)	Mean (3)	SD (4)	Mean (5)	SD (6)
EGRA						
Letter Name Knowledge (Letters per Minute)	476	15.3%	1.180	4.424	4.857	9.349
Initial Sound Identification (Sounds Identified)	477	2.9%	0.161	1.028	0.455	2.011
Familiar Word Reading (Words per Minute)	476	1.3%	0.168	1.617	0.165	2.588
Invented Word Reading (Words per Minute)	474	0.6%	0.084	1.191	0.275	2.309
Oral Reading Fluency (Words per Minute)	474	1.9%	0.508	4.537	0.102	5.012
Reading Comprehension (Questions Correct)	477	30.0%	0.327	0.559	-0.111	0.703
English Oral Assessment						
Test 1 (Vocabulary)	477	58.5%	1.774	1.993	0.275	2.089
Test 1 (Count)	477	32.9%	0.501	0.771	-0.208	0.813
Test 2a (Vocabulary)	477	36.9%	0.669	1.008	-0.168	1.068
Test 2a (Phrase Structure)	477	36.3%	0.801	1.169	0.006	1.343
Test 2b (Vocabulary)	477	54.9%	1.400	1.655	0.426	2.079
Test 2b (Phrase Structure)	477	48.4%	1.520	1.892	0.572	2.512
Test 3 (Vocabulary, Expressive - Objects)	477	67.1%	2.365	2.436	-0.038	2.490
Test 3 (Vocabulary, Expressive - People)	477	52.2%	1.505	1.789	0.080	2.177
Writing Test						
African Name (Surname) Spelling & Capitalization	477	20.1%	0.201	0.401	0.392	0.654
English Name (Given name) Spelling & Capitalization	477	14.5%	0.145	0.352	0.193	0.499
Ideas	477	0.6%	0.006	0.079	0.135	0.360
Organization	477	0.2%	0.002	0.046	0.284	0.589
Voice	477	0.0%	0.000	0.000	0.164	0.393
Word Choice	477	6.9%	0.069	0.254	0.099	0.374
Sentence Fluency	477	0.6%	0.006	0.079	0.261	0.584
Conventions	477	0.0%	0.000	0.000	0.116	0.339

Notes: Statistics are for the 477 control-group members of the Longitudinal Sample, which includes students who were tested at baseline as well as endline. Change from Baseline to Endline is the student's endline score on the component minus his or her baseline score.

Table 4: Program Impacts on Early Grade Reading Assessment Scores (in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	PCA EGRA Score Index†	Letter Name Knowledge	Initial Sound Recogniton	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Mango Tree-Administered Program	0.634***	1.014***	0.647***	0.374***	0.215**	0.476***	0.445***
	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.129)	(0.113)
Government-Administered Program	0.133	0.407**	0.076	-0.002	0.031	0.071	0.045
	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
Number of Students	1438	1475	1481	1471	1467	1450	1481
Number of Schools	38	38	38	38	38	38	38
Adjusted R-Squared	0.153	0.219	0.103	0.067	0.076	0.075	0.058
Control Group Mean§	0.002	5.977	0.616	0.335	0.360	0.615	0.216
Control Group SD§	1.005	9.374	1.920	2.209	2.770	4.176	0.437

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

† PCA EGRA Score Index is constructed by normalizing each of the 6 test modules (columns 2 through 7) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated effects are comparable but slightly larger for an alternative index that uses the unweighted mean across test modules instead.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the index (column 1), where they are the normalized values.

Table 5: Program Impacts on Oral English Test Scores & English Word Recognition (in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Oral English Score Index†	Test 1 (Vocab.)	Test 1 (Count)	Test 2a (Vocab.)	Test 2a (Phrase Structure)	Test 2b (Vocab.)	Test 2b (Phrase Structure)	Test 3 (Vocab., Expressive - Objects)	Test 3 (Vocab., Expressive - People)	Recognition of Printed English Words‡
Mango Tree-Administered Program	0.141	0.157	-0.118	-0.034	0.045	0.025	-0.114	0.306***	0.295**	-0.290**
	(0.100)	(0.099)	(0.097)	(0.095)	(0.114)	(0.100)	(0.113)	(0.105)	(0.117)	(0.135)
Government-Administered Program	-0.089	0.001	-0.115	-0.020	-0.113	-0.154	-0.213*	-0.023	-0.099	-0.209
	(0.091)	(0.082)	(0.091)	(0.103)	(0.092)	(0.095)	(0.119)	(0.095)	(0.086)	(0.140)
Number of Students	1481	1481	1481	1481	1481	1481	1481	1481	1481	1481
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.346	0.164	0.163	0.205	0.186	0.279	0.0920	0.238	0.188	0.274
Control Group Mean§	0	2.048	0.294	0.501	0.807	1.826	2.092	2.327	1.585	1.792
Control Group SD§	1	1.888	0.620	0.911	1.209	1.928	2.217	2.133	1.839	4.184

Notes: Longitudinal sample includes 1,478 students 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 Recognition of Printed English Words (column 10), which was not administered at baseline. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 8 test modules (columns 2 through 9) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated are comparable but slightly larger in magnitude for an alternative index that uses the unweighted mean across test modules instead.

‡ Recognition of Printed English Words is not part of the Oral English examination, but it is a skill that is commonly practiced in *status quo* (i.e. control) schools in the Lango sub-Region. This involves reading a set of 18 printed words from a piece of paper. It is not included in the computation of the overall PCA index in column 1.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indi

Table 6: Program Impacts on Writing Test Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	African Name (Surname) Writing	English Name (Given Name) Writing	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Mango Tree-Administered Program	0.422***	0.922***	1.312***	0.163	0.441**	0.152	0.175	0.383*	0.221	0.139
	(0.146)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
Government-Administered Program	-0.172	0.435***	0.450***	-0.274*	-0.316*	-0.313**	-0.262**	-0.330*	-0.253	-0.330**
	(0.125)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
Number of Students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.356	0.240	0.236	0.174	0.304	0.177	0.200	0.302	0.164	0.171
Control Group Mean§	0	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control Group SD§	1	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students 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. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 9 test modules (columns 2 through 10) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

Table 7: Program Impacts on Student Aspirations, Preferences, and Effort from Endline Survey

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variable	Pupil Thinks He/She will Pass PLE at End of P7	Preference for School over Other Activities†	Prefers Literacy to Math Class	Wants a Career as a Doctor/Nurse	Wants a Career as a Headmaster/ Teacher	Practices Writing at Home	Thinks He/She is a Good Student	Perceived Rank in Class‡	Career Ambition Rating††
Units	Percentage Points	Control Group SD	Percentage Points	Percentage Points	Percentage Points	Percentage Points	Percentage Points	Control Group SD	Control Group SD
Mango Tree-Administered Program	0.022**	-0.114	-0.000	-0.078**	0.071***	0.006	0.002	0.148**	-0.059
	(0.009)	(0.112)	(0.023)	(0.033)	(0.023)	(0.025)	(0.013)	(0.063)	(0.068)
Government-Administered Program	0.015*	-0.097	-0.021	-0.030	0.035	-0.002	0.006	0.018	-0.085
	(0.009)	(0.087)	(0.021)	(0.027)	(0.024)	(0.020)	(0.015)	(0.076)	(0.056)
Number of Students	1330	1470	1457	1427	1427	1420	1371	1333	1417
Number of Schools	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	-0.002	0.003	0.005	0.024	0.008	0.009	0.004	0.027	0.026
Control Group Mean§	0.947	4.614	0.544	0.396	0.154	0.900	0.971	2.245	2.837
Control Group SD§	0.225	0.657	0.499	0.490	0.361	0.300	0.169	0.666	0.886

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† Preference for School over Other Activities is a 5-point scale based on a list of questions that compared school activities to other activities, capturing the number for which the student expressed a preference for school (and omitting those where she provided no response or could not answer).

‡ Perceived Rank in Class is a 1-3 scale, with 1 being the bottom of the class, 2 being the middle of the class, and 3 being the top of the class.

†† Career Amibition Rating is a subjective 1-5 scale where 1 is the least ambitious and 5 is the most ambitious; the ratings for each career were done by an evaluator who was blinded to the treatment status of the pupils.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations.

Table 8: Classroom Observations – Teacher Behavior

	(1) Refers to Teacher's Guide	(2) % time Speaking Leblango	(3) Observe/ Record Performance	(4) Moves Freely	(5) Remains at Front of Class	(6) Encourages Participation	(7) Brings Pupils back on Task	(8) Ignores Off-Task Students
Mango Tree-Administered Program	0.035	11.513***	0.047	0.087	-0.121**	-0.004	0.007	0.056**
	(0.041)	(3.524)	(0.052)	(0.067)	(0.053)	(0.018)	(0.038)	(0.027)
Government-Administered Program	0.041	8.907**	-0.025	-0.007	-0.048	-0.001	-0.070	0.062**
	(0.036)	(3.592)	(0.048)	(0.045)	(0.044)	(0.018)	(0.043)	(0.025)
Number of Observations	441	438	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.166	0.121	0.256	0.032	0.061	-0.004	0.061	0.006
Control Group Mean[§]	0.802	67.210	0.237	0.733	0.237	0.962	0.870	0.031

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 9: Classroom Observations – Student Behavior While Reading

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Students are Reading:								
	Sounds	Letters	Words	Sentences	On Board	Reading From Primer	Reading From Reader	% of Reading Done in Leblango	Minutes Spent on Reading
Mango Tree-Administered Program	0.113***	-0.004	0.050	0.124***	-0.053	0.121***	0.064***	0.219***	0.669***
	(0.034)	(0.043)	(0.043)	(0.042)	(0.043)	(0.025)	(0.023)	(0.050)	(0.242)
Government-Administered Program	0.067**	0.054	-0.025	0.019	0.021	0.069***	0.031	0.165***	0.523**
	(0.028)	(0.044)	(0.045)	(0.051)	(0.039)	(0.022)	(0.020)	(0.051)	(0.212)
Number of Observations	441	441	441	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.015	0.007	0.047	0.018	0.039	0.041	0.165	0.039	0.083
Control Group Mean[§]	0.061	0.206	0.649	0.282	0.672	0.023	0.038	0.466	3.687

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 10: Classroom Observations – Student Behavior While Writing

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Students are Writing:									
	Pictures	Letters	Words	Sentences	Their Names	Air Writing	Copying Text from Board	Writing Own Text	% of Writing Done in Leblango	Minutes Spent on Writing
Mango Tree-Administered Program	0.076**	-0.024	0.044	0.023	0.059**	0.019	-0.024	0.094***	0.108**	0.199
	(0.033)	(0.035)	(0.028)	(0.023)	(0.028)	(0.030)	(0.036)	(0.028)	(0.047)	(0.253)
Government-Administered Program	0.034	0.042	0.059*	-0.028	0.006	0.063**	-0.017	0.019	0.115***	0.294
	(0.031)	(0.034)	(0.029)	(0.017)	(0.023)	(0.029)	(0.038)	(0.022)	(0.041)	(0.227)
Number of Observations	441	441	441	441	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.012	-0.012	0.007	0.004	0.055	0.007	0.024	0.034	-0.002	0.001
Control Group Mean[§]	0.069	0.115	0.084	0.038	0.046	0.076	0.130	0.061	0.168	1.237

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 11: Classroom Observations – Student Behavior While Writing

	(1)	(2)	(3)	(4)	(5)	(6)
	Students are Speaking and Listening:				% of Speaking and Listening Done in Leblango	Minutes Spent on Speaking and Listening
	To Partner	To Small Group	To Whole Class	To Teacher		
Mango Tree-Administered Program	-0.028	0.050*	-0.041	-0.064*	0.080**	-0.786**
	(0.045)	(0.029)	(0.043)	(0.036)	(0.035)	(0.325)
Government-Administered Program	-0.014	0.066**	0.006	-0.094**	0.067*	-0.330
	(0.036)	(0.031)	(0.037)	(0.036)	(0.033)	(0.540)
Number of Observations	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38
Adjusted R-Squared	0.276	0.025	0.140	0.103	0.068	0.062
Control Group Mean§	0.221	0.038	0.748	0.947	0.802	4.916

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 12: Attendance and Enrollment

	(1)	(2)	(3)	(4)	(5)	(6)
	Pupil Attendance:				Pupil Enrollment	Teacher Surveys
	Present for Visit 1	Present for Visit 2	Present for Visit 3	Average across all 3 visits	Total Enrollment at Endline	Reports Having Missed School in Past Month
Mango Tree-Administered Program	0.105**	0.020	0.026	0.050*	2.364	-0.067
	(0.044)	(0.031)	(0.035)	(0.029)	(25.008)	(0.167)
Government-Administered Program	0.019	-0.062**	-0.080**	-0.041	2.797	0.109
	(0.046)	(0.026)	(0.034)	(0.031)	(27.545)	(0.169)
Number of Observations	5334	5334	5334	5334	38	71
Number of Schools	38	38	38	38	38	37
Adjusted R-Squared	0.026	0.023	0.032	0.038	0.017	-0.025
Control Group Mean[§]	0.459	0.406	0.405	0.423	233.3	0.348

Notes: All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.
§ Control Group Mean is computed using the endline data for control-group observations in the estimation sample.

Table 14: Responses to Teacher Survey by Study Arm

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Weekly Hours Spent on:													
	Teaching	Prep. for Class	Helping Students Outside Class	Taught Literacy this Year	# Parents Met with This Year	Would Choose to Teach if Could Restart Career	Teacher's Fault if Students Don't Learn	Satisfied with P2/P3 Reading at This School	Rating of Own Teaching Compared to Rest of School (1-3)	% of Pupils Teacher thinks will Pass PLE	Attended Any Training this Year	Days of Training Attended This Year	Went to NGO-Provided Training	Went to Other Training
Mango Tree-Administered Program	1.904	-0.623	2.042*	0.176*	61.288	0.199	-0.342**	-0.539***	0.015	0.003	0.319***	3.147*	0.567***	-0.255**
	(2.206)	(2.643)	(1.126)	(0.097)	(45.124)	(0.124)	(0.160)	(0.098)	(0.150)	(0.108)	(0.105)	(1.558)	(0.115)	(0.099)
Government-Administered Program	1.808	1.902	0.547	0.313***	35.918	0.200*	-0.034	-0.434***	-0.324**	-0.123	0.295***	2.348	0.470***	-0.169
	(2.494)	(2.851)	(0.970)	(0.095)	(34.203)	(0.105)	(0.159)	(0.094)	(0.150)	(0.094)	(0.105)	(2.043)	(0.121)	(0.110)
Number of Observations	73	72	69	73	67	70	72	71	71	73	73	70	73	73
Number of Schools	38	38	38	38	36	37	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.101	0.219	-0.0480	0.203	0.0810	0.0370	0.150	0.245	0.146	0.197	0.131	0.0940	0.326	0.0940
Control Group Mean§	14.55	9.601	1.765	0.609	37.86	0.565	0.739	0.727	2.545	0.498	0.652	4.957	0.435	0.348
Control Group SD§	8.780	10.67	2.221	0.499	46.94	0.507	0.449	0.456	0.510	0.291	0.487	6.852	0.507	0.487

Notes: All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample.

Table 14: Cost-Effectiveness Calculations

	Program Administered By:	
	Mango Tree	**Government**
Cost per student	$13.98	$4.47
Letter Name Knowledge		
Effect Size (SDs)	1.01	0.41
Cost per student/0.2 SDs	$2.76	$2.20
SDs per dollar	0.07	0.09
PCA EGRA Index		
Effect Size (SDs)	0.63	0.13
Cost per student/0.2 SDs	$4.41	$6.72
SDs per dollar	0.05	0.03
PCA Writing Test Index		
Effect Size (SDs)	0.42	-0.17
Cost per student/0.2 SDs	$6.63	N/A
SDs per dollar	0.03	-0.04

Appendix A: Intervention Inputs

The Mango Tree and Government Administered Programs differ in terms of the materials, training, and other support provided to schools; we specify the differences for each below, and also show them in Table 1.

Materials

The NULP provides the following materials to each MT and CCT school:

- One Leblango Teacher's Guide for each teacher
- Three term-specific Leblango primers for each student (up to 200 students per class)
- Three term-specific Leblango readers for each student (up to 200 students per class)
- One English Teacher's Guide for each P1-P3 teacher
- Three term-specific English primers for each student (up to 200 students per class)

In addition, the MT Program provides additional materials to each school:

- One slate for each student (up to 200 students per class)
- Two wall clocks per school

Teacher Training

The NULP's teacher training comprises the following:

- One residential five-day training in the Leblango orthography for P1-P3 teachers in December the year before they enter the program (MT Program only)
- Three trainings in literacy methods for P1-P3 teachers during the school holidays each year
    - MT Program: residential trainings held in the district capital, conducted by experienced MT staff
    - CCT Program: non-residential trainings held at the CCs, conducted by CCTs. To facilitate these trainings, Mango Tree CCTs with instructional videos to learn which they play on solar-powered, portable DVD players. The videos also provide examples of instructional practice in real-life classrooms, as well as provide a possible inexpensive alternative to residential training models.
- Special field monitoring and support supervision visits to schools
    - MT Program: 3 times per term by project staff, 2 times per term for CCTs
    - CCT Program: 2 times per term for CCTs

Other Support

- Parent Interaction. Schools in both the MT Program and CCT Program hold a parent meeting each term. Each meeting has specific content designed by Mango Tree as well as time for other school-related issues to be addressed. These meetings are conducted by the field officers for the MT Program schools and the CCTs for the CCT Program schools. The term 1 meeting focuses on answering parents' questions about literacy and the NULP. It also introduces a specialized report card, which differs from the ones ordinarily used by school, that the NULP uses to provide parents with feedback on their children's performance. The term 2 meeting allows parents to observe classes in session and trains parents in the Parent Assessment Tool. Modeled after one developed in India by Pratham and also used by UWEZO in East Africa, the tool a

simple way for parents to assess their students in basic reading skills.[1] At the term 3 meetings, students demonstrate what they've learned during the school year for their parents and are awarded prizes for a variety of literacy and other academic achievements.

- Monthly Radio Program. Mango Tree sponsors a one-hour monthly radio program (supported by SMS messages and surveys to engage listeners in feedback) that broadcasts literacy and local language education topics to parents, teachers and communities in the Lango Sub-region. This program is available to students, teachers, and parents in all three study arms, and thus we cannot analyze its effects in this study.

- Take a Book Home Activity (MT Program only). Beginning near the end of the first term, children take home books each week that they are expected to read with their parents and other family members. Teachers are given a simple recording sheet to track the movement of books.

[1] The tool has 4 parts: 1) letter name knowledge; 2) familiar word reading; 3) reading fluency test; and 4) reading comprehension test.

Appendix B: Robustness Checks

B.1 Effect of NULP on Exam Scores without Controlling for Baseline Scores

Our preferred specification for analyzing the effect of the NULP on exam scores controls for the pupil's baseline score on the test component in question, or when analyzing the effect on the combined exam score indices, controls for the pupil's baseline score on the index. In this section, we show that our results are qualitatively and numerically robust to the exclusion of those controls from our regressions. In this section we replicate Tables 4-6, but instead of estimating equation (1) we estimate:

$$(5) \qquad y_{is}=\beta_0+\beta_1 MTSchool_s+\beta_2 CCTSchool_s+\mathbf{L}_s'\boldsymbol{\gamma}+\varepsilon_{is}$$

Here i indexes students and s indexes schools. y_{is} is a student's endline score on a particular exam or exam component. $\boldsymbol{L}_s$ is a vector of indicator variables for the stratification group that a school was in for the public lottery that assigned schools to study arms. This specification differs from (1) solely in that it omits $y^{baseline}_{is}$, the student's baseline score on the test component, from the right-hand side.

The results are presented in Appendix Tables B1 to B3, which mirror tables 4 to 6 in the main text. The point estimates and standard errors are nearly unaffected by the exclusion of the controls. For the EGRA (Appendix Table B1), including the regression without baseline test score results yields to slightly larger effect sizes for the Mango Tree-Administered Program and slightly smaller effect sizes for the Government-Administered Program.

For the Oral English Test (Appendix Table B2)[2] and the Writing Test (Appendix Table B3)[3], omitting the baseline test score controls leads to marginally smaller estimates of the gains for students in the Mango Tree-Administered variant of the program, and marginally larger estimated losses for students in the Government-Administered version. The exception is the two name-writing components of the Writing Test, for which the students receiving the Government-Administered version of the program showed gains rather than losses. For African Name (Surname) Writing, the estimated effect of the Government-Administered program differs only in the third decimal place. For English Name (Given Name) Writing, the estimated effect is somewhat smaller without controlling for baseline performance.

None of the differences affect the statistical significance of any of the point estimates, nor do they change any of the conclusions we draw in the main text.

---

[2] Note that Column 10 is identical between Table 4 and Appendix Table A2; no controls were included for this column in Table 4 because this test, which is not a component of the Oral English Examination, was not conducted at baseline.

[3] Column 10 is identical between Table 5 and Appendix Table A3 because Presentation was not one of the scored categories at baseline. Columns 6 (Voice) and 9 (Conventions) are also identical because no pupils received any points for those categories at baseline, so the controls were dropped due to collinearity with the constant term.

Appendix Table B1: Program Impacts on Early Grade Reading Assessment Scores, without Controlling for Baseline Scores (in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	PCA EGRA Score Index†	Letter Name Knowledge	Initial Sound Recogniton	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Mango Tree-Administered Program	0.654***	1.043***	0.649***	0.382***	0.233**	0.484***	0.449***
	(0.127)	(0.163)	(0.129)	(0.0909)	(0.0967)	(0.121)	(0.110)
Government-Administered Program	0.110	0.418**	0.0639	-0.0116	0.0206	0.0581	0.0337
	(0.102)	(0.181)	(0.0956)	(0.0742)	(0.0692)	(0.0807)	(0.0837)
Number of Students	1460	1476	1481	1474	1471	1467	1481
Number of Schools	38	38	38	38	38	38	38
Adjusted R-Squared	0.118	0.175	0.0965	0.0559	0.0367	0.0629	0.0509
Control Group Mean§	0.000	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 who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 6 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated effects are comparable but slightly smaller for an alternative index that uses the unweighted mean across test modules instead.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the index, where they are the normalized values.

Appendix Table B2: Program Impacts on Oral English Test Scores & English Word Recognition, without Controlling for Baseline Scores (in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Oral English Score Index†	Test 1 (Vocab.)	Test 1 (Count)	Test 2a (Vocab.)	Test 2a (Phrase Structure)	Test 2b (Vocab.)	Test 2b (Phrase Structure)	Test 3 (Vocab., Expressive - Objects)	Test 3 (Vocab., Expressive - People)	Recognition of Printed English Words‡
Mango Tree-Administered Program	0.0677	0.122	-0.133	-0.0723	0.0155	-0.0138	-0.120	0.275**	0.291**	-0.290**
	(0.123)	(0.108)	(0.0936)	(0.106)	(0.131)	(0.112)	(0.117)	(0.117)	(0.119)	(0.135)
Government-Administered Program	-0.133	-0.0194	-0.124	-0.0397	-0.130	-0.165	-0.223*	-0.0405	-0.106	-0.209
	(0.102)	(0.0864)	(0.0878)	(0.108)	(0.0988)	(0.102)	(0.120)	(0.0985)	(0.0883)	(0.140)
Number of Students	1481	1481	1481	1481	1481	1481	1481	1481	1481	1481
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.319	0.155	0.162	0.199	0.178	0.268	0.0900	0.230	0.183	0.274
Control Group Mean§	0.000	2.048	0.294	0.501	0.807	1.826	2.092	2.327	1.585	1.792
Control Group SD§	1.000	1.888	0.620	0.911	1.209	1.928	2.217	2.133	1.839	4.184

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 8 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated are comparable but slightly larger in magnitude for an alternative index that uses the unweighted mean across test modules instead.

‡ Recognition of Printed English Words is not part of the Oral English examination, but it is a skill that is commonly practiced in status quo (i.e. control) schools in the Lango sub-Region. This involves reading a set of 18 printed words from a piece of paper. It is not included in the computation of the overall PCA index in column 1.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

Appendix Table B3: Program Impacts on Writing Test Scores, without Controlling for Baseline Scores (in SDs of the Control Group Endline Score Distribution)

	(1) PCA Writing Score Index†	(2) African Name (Surname) Writing	(3) English Name (Given Name) Writing	(4) Ideas	(5) Organization	(6) Voice	(7) Word Choice	(8) Sentence Fluency	(9) Conventions	(10) Presentation
Mango Tree-Administered Program	0.399**	1.015***	1.230***	0.147	0.442**	0.152	0.128	0.377*	0.221	0.139
	(0.186)	(0.116)	(0.148)	(0.178)	(0.207)	(0.156)	(0.178)	(0.210)	(0.173)	(0.150)
Government-Administered Program	-0.232	0.437***	0.393**	-0.288*	-0.317*	-0.313**	-0.308**	-0.334*	-0.253	-0.330**
	(0.163)	(0.127)	(0.152)	(0.150)	(0.178)	(0.134)	(0.151)	(0.179)	(0.156)	(0.129)
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	0.265	0.193	0.217	0.161	0.304	0.177	0.165	0.300	0.164	0.171
Control Group Mean§	0	0.593	0.350	1.141	1.286	1.164	1.166	1.267	1.116	1.175
Control Group SD§	1	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 who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 11 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

B.2 Effect of NULP on Writing Scores, Excluding Stratification Cell of School that Completed Writing Test in English

Students from one of the 12 control schools were mistakenly asked to complete their writing tests in English. The name-writing components of the test were unchanged, and the tests were scored using the exact same rubric as the Leblango writing test. However, there is still the potential concern that the tests from this school may not be comparable to those from the other 37 schools. To address this possibility we re-estimate equation (1) for the writing test, excluding the stratification cell for the school that completed the test in English. This stratification cell includes one school from each of the other two study arms as well, so dropping the cell yields a reduced sample of 35 schools. Since the random assignment of schools to study arms was conducted within stratification cells, the exogeneity assumption that *MTSchool* and *GovtSchool* are independent of ε_{is} will also hold for this reduced sample. In the presence of treatment effect heterogeneity, however, we would not expect this sample to produce identical treatment effect estimates even if there were no issues with the control school's tests.

Appendix Table B4 shows the estimated effects of the two program variants on test scores using the reduced sample described above. Excluding this cell changes the magnitude of the estimated effects, but does not change their sign or affect our interpretation of them. The estimated gains from the Mango Tree-administered version of the program are similar but somewhat larger; the combined PCA index shows a 50% larger increase using the reduced sample. For the Government-administered program, the combined index shows a fairly precise zero change. The improvements in name-writing are similar to the full sample, while the declines in the other exam components are smaller. Nevertheless, two of the seven writing components show statistically-significant decreases in performance, as compared with three for the full sample. Overall, the results are not particularly sensitive to the inclusion of this stratification cell.

Appendix Table B4: Program Impacts on Writing Test Scores, Excluding Stratification Cell for School that Completed Exam in English (in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	African Name (Surname) Writing	English Name (Given Name) Writing	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Mango Tree-Administered Program	0.594***	0.933***	1.364***	0.372***	0.701***	0.350***	0.351***	0.638***	0.435***	0.328***
	(0.107)	(0.117)	(0.150)	(0.109)	(0.129)	(0.091)	(0.114)	(0.130)	(0.110)	(0.088)
Government-Administered Program	-0.010	0.473***	0.527***	-0.093	-0.079	-0.130**	-0.107	-0.093	-0.050	-0.155**
	(0.075)	(0.125)	(0.149)	(0.078)	(0.088)	(0.060)	(0.078)	(0.085)	(0.082)	(0.060)
Number of Students	1262	1336	1263	1361	1361	1360	1360	1361	1361	1361
Number of Schools	35	35	35	35	35	35	35	35	35	35
Adjusted R-Squared	0.323	0.234	0.241	0.153	0.319	0.165	0.151	0.302	0.146	0.158
Control Group Mean§	-0.261	0.527	0.274	0.0610	0.131	0.0840	0.0750	0.108	0.0370	0.0980
Control Group SD§	0.585	0.671	0.486	0.239	0.338	0.278	0.264	0.310	0.190	0.298

Notes: Sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators as well as baseline values of the outcome variable, except for "Presentation" (column 10) which was not included in the baseline scores. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 11 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.