Reading for Life: Lasting Impacts of a Literacy Intervention in Uganda

Julie Buhl-Wiggers, Jason T. Kerwin,
Ricardo Montero de la Piedra, Jeffrey Smith, and Rebecca Thornton*

January 17, 2024

Abstract

Despite many examples of educational programs that are effective in the short run, evidence of programs that remain effective in the long-run remains scarce. We study the Northern Uganda Literacy Project (NULP)—a three year early grade reading intervention that resulted in large short-term impacts (1.2 SD in local language reading and 0.5 SD in English reading). We follow students 8-9 years after the program began (and 5-6 years after it ended) and find 58% of the effect remains in local language and 100% remains in English. These effects represent 4.4 extra years of local-language learning and 1.6 extra years of English learning compared to the control group. We find no spillover effects on math or sexual behavior. The control group exhibits dismal grade progression and retention both before and during the COVID-19 pandemic. While the NULP had no impact on attendance or remaining in school, it modestly improved grade progression.

JEL Codes: I2, O1

Keywords: Economics of Education, Human Capital, Development Economics, Long-run, Treatment Effects, Fade-out

*Buhl-Wiggers: Department of Economics, Copenhagen Business School (jubu.eco@cbs.dk); Kerwin: Department of Applied Economics, University of Minnesota and J-PAL (jkerwin@umn.edu); Montero de la Piedra: Department of Applied Economics, University of Minnesota (monte290@umn.edu); Smith: Department of Economics, University of Wisconsin (econjeff@ssc.wisc.edu) Thornton: Department of Economics, Baylor University (rebecca_thornton@baylor.edu). The randomized evaluation of the Northern Uganda Literacy Project would not have been possible without the collaboration of Victoria Brown and Ichuli Institute, Katherine Pollman, Deborah Amuka and other Mango Tree Educational Enterprises staff. We thank Dave Evans, Deion Filmer, Dan Gilligan, Sarah Kabay, Ibrahim Kasirye, and seminar audiences at CESifo, RISE, and the World Bank for helpful comments and suggestions. We are grateful for funding from DFID/ESRC Raising Learning Outcomes Grant ES/M004996/2, Wellspring, and the International Growth Centre.

1 Introduction

Learning gains rarely last. Although many education interventions have been shown to improve learning in the short run, 80 percent of those gains disappear within just four months (Hattie, Biggs, and Purdie 1996). This pattern of "fade-out" is well-documented in rich country education systems, occurring in domains ranging from reading ability (Bus and IJzendoorn 1999) to cognitive skills (Protzko 2015; Takacs and Kassai 2019). Bailey et al. (2020) conclude that fade-out widespread and not just the result of measurement error.

Little is even known about the longer-term impacts of increases in learning. The literature on fade-out implies that they will dissipate, with the treatment group falling back to its previous trend or the control group catching up. But there is evidence that some interventions do have lasting benefits, with effects re-emerging later (Bailey et al. 2020). This evidence is strongest for early-childhood development programs like the Perry Preschool Project (Heckman, Pinto, and Savelyev 2013). There is also evidence that better teachers can have lasting impacts on student outcomes (Chetty et al. 2011). What is much less clear is whether interventions to improve learning within existing school systems can also achieve lasting gains.

To answer this question, we study the long-run impacts of one of the most effective education programs in the world. The Northern Uganda Literacy Project (NULP) targeted students in government schools in grades 1-3. It was based on a structured pedagogy approach, with teachers instructed in exactly how to teach each lesson. This approach was implemented through teacher guides with daily lesson plans and textbooks that were linked to the guides. Teachers received intensive training in how to implement the program, and support visits from the trainers to give them feedback on their teaching. Our data comes from a randomized trial that assigned 128 schools to a control group, the original NULP, or a reduced-cost version that eliminated some of the costlier inputs. Previous research shows that the NULP increases reading scores by well over 1 SD on average (Buhl-Wiggers et al. 2022). These short-run impacts are very large, at the 99th percentile of all interventions that have been assessed via randomized trials in the developing world (Evans and Yuan 2022).

We are able to locate 93 percent of the original sample of students from the study between 8 and 9 years after the program began.[1] Since the program ran for three years, from 2014 to 2016, this means that our followup data comes from 5 and 6 years after the intervention ended. It also means that we successfully tracked our sample through the COVID-19 pandemic, during which Uganda closed its schools longer than any other country (for nearly

[1] Due to a limited budget for data collection, we did not complete exams and surveys for all 93 percent of students; our completed data collection rate was 75 percent.

two years). We achieved this high tracking rate by targeting students based on exogenous, pre-treatment characteristics that predicted low attrition, and via engagement and partnership with the communities we are studying. Our analyses are all based on a pre-analysis plan unless otherwise specified, and our results are all robust to adjustment for multiple hypothesis testing.

The vast majority of the program's impacts are still present between 5 and 6 years after the intervention ends. Specifcally, children in the treatment-group schools are still ahead in local-language reading by 0.71 SDs, equivalent to 4.4 school years of learning under the status quo, which is 58 percent of the initial impact of the program. In English-language reading there is no fade-out at all: the initial impact is 0.51 SDs, and the effect 5-6 years later is 0.55 SDs. This long-run impact is equivalent to the gains from 1.6 years of schooling in the control group. The NULP focuses on mother-tongue-first instruction, and teaches students to read in English only in grades 2 and 3. Thus the stronger persistence of impacts in English are suggestive of positive spillovers onto second-language reading skills. The overall school system transitions to English-language instruction by grade 5, and so the lasting treatment effects on English may mean that the treatment group has an ongoing advantage at school.

Consistent with the NULP leading to persistent educational advantanges, we find that the program mitigated the grade delay that plagues the Ugandan school system. If the children in our study had progressed "on-track" (e.g., moving up one grade each year), they would have reached secondary school (e.g., grade eight) by 2021. Yet, we find that essentially none of them have reached grade 8, and just 3% of the control group has even reached grade 7. Indeed, the vast majority of control-group children are more than three grades behind their expected level. The onset of COVID-19 in 2020 only somewhat exacerbated this pre-existing pattern. The NULP mitigated this negative trajectory: students in the treatment group experienced about 10% less delay in grade progression than the control group, a statistically significant advantage. They were substantially more likely to be ahead of typical grade progression, and this gap opened up in grade 4, just after the intervention ended. However, they were no more likely to be enrolled in school at all, nor to reach secondary school (since no one did).

In contrast to these encouraging effects on reading and grade progression, we find no spillover effects onto math scores nor onto other life outcomes. For math, we find null results just after the program ended, and no change in them 5-6 years later. However, math requires the less in the way of reading skills than science or social studies, and we lack data on these other subjects. We also measure impacts on working outside the home and sexual activity, finding null results on both. It may be to early to see effects on these outcomes, particularly on sexual activity; 54 percent of the sample works outside the home, but just 11 percent has

ever had sex.

One of this paper's main contributions is to build on the limited evidence on the persistence of education program impacts in developing countries. Most existing evidence on this topic comes from studies of conditional cash transfers (Bouguen et al. 2019). These studies find consistently positive long-run effects on schooling, and some evidence of impacts on learning and labor market outcomes (Millán et al. 2019). Other studies focus on interventions providing school uniforms or scholarships. In Kenya, Evans and Ngatia (2021) do not find long-run effects of providing school uniforms in primary schools—even though school absenteeism decreased in the short run. In Colombia and Ghana, scholarships for secondary education increased school attainment and labor market participation, and also decreased teen fertility (Bettinger et al. 2019; Duflo, Dupas, and Kremer 2021).[2]

While most of the existing knowledge base derives from interventions that increase the *quantity* of schooling through school attendance, our study evaluates an intervention aimed at improving the *quality* of schooling and finds relatively long-lasting effects on learning. Most of the more recent interventions that have been implemented and studied in the recent two decades have focused on improvements on this margin (Evans and Yuan 2022). But as Hares, Rossiter, and Sandefur (2023) highlights, there is limited long-run followup evidence on this body of work. The only other study along these lines that we know of is Stern et al. (2023), which follows participants in a reading program in South Africa 4 years after the intervention ends and also finds that some of the effects persist. Relative to their study, we are able to follow participants in an intervention with far larger impacts, and the effects of the NULP are more persistent. This means that our null results for math and life outcomes are more informative. It also suggests a potential explanation for the greater persistence of our impacts, via the "Matthew Effect". A complementary finding of the two studies is that both find evidence of improvements in grade progression.

Our results also shed light on the viability of "reading to learn": the idea that boosting reading skills will lead to more success in school. This concept has been promoted by non-profits (Michie 2023; Center for Public Education 2015) and education schools (Harvard Graduate School of Education 2016). "Reading to Learn" is also the name of a major education program that has been implemented around the world. Our study provides evidence on whether it works. We see promising evidence that foundational reading skills lead to greater success later in school, but on the other hand no impacts on math scores. The effects on

[2] A related strand of the literature evaluates school health or early-childhood development programs (e.g. Baird et al. 2016; Gertler et al. 2014; Barham et al. 2023). Some of these interventions show effects on school participation as well as improved life outcomes as much as 20 years later (Baird et al. 2016; Hamory et al. 2021; Gertler et al. 2014). However, only interventions aimed at very early ages seem to affect learning (Ozier 2018).

other subjects remain an open question.

Finally, our program sheds light on the effectiveness of education interventions as a way to ameliorate the harms of pandemics and school closures. Uganda had the longest COVID-19 school closures in the world, and there was great concern about how this would affect children (Athumai 2022). Evidence from the Ebola outbreak in Sierra Leone suggests that school closures led to earlier sexual debut for girls, with ensuing effects on pregnancy and school dropout (Bandiera et al. 2023). Moreover, they find that a "safe space" intervention helped to mitigate this problem. We find differing results: during Uganda's COVID-19 school closures, self-reported sexual activity remained low and improved reading ability and grade progression did not change that. At the same time, we do find that the NULP mitigated the grade retention that was partly attributable to the school closures, which suggests that investments in foundational literacy could be valuable inoculations against the harms to learning from future pandemics.

This paper proceeds as follows: First we set the stage and describe the NULP program in Section 2. In Section 3 we describe the original evaluation of the NULP as well as the sampling frame and long-term follow-up data collection. In Section 4 we describe our identification strategy. In Section 5 we present the results in three stages. First, we present how educational outcomes progressed from 2014 to 2021 in absence of the program. Sencond, we present the impacts of the NULP immediately after the program ended in 2016. Finally, we present the long term impacts measured in 2021—five years after the program ended. Section 6 concludes.

2 Setting

2.1 Primary Education in Uganda

Primary education in Uganda consists of seven years of schooling (P1 to P7) and the official school starting age is six years. Since 1997, primary school has officially been free of charge, however, as resources are scarce many schools still depend on contributions from parents, thus *de facto* school fees are common and students whose parents are not able to meet these contributions are often sent home. The reform of 1997 was successful in enrolling children into school, especially at earlier grades; the net primary school enrollment rate is above 90% for both boys and girls (World Bank 2020). Yet, the large influx of children and limited resources has created raising concerns about diminishing school quality. Dropout and delayed progression through school remains an issue for both boys and girls.

The Ugandan national policy is to conduct all instruction in grades one to three (P1

to P3) using the local mother-tongue language. Grade four (P4) is a transition year, and grades five through seven (P5-P7) are taught in English. In grades one to three, students are taught reading and writing for an hour each day in their local language, and receive English instruction for a half an hour. The government provides primary school teachers with a variety of materials such as teachers' guides, resource books, and student learning materials. Often, however the delivery and use of these materials are inadequate.

Teachers in Ugandan primary schools receive their basic teacher training at a Primary Teacher College (PTC). Once hired, teachers receive additional training and continuous professional development through the Teacher Development and Management System (TDMS). Under this system, Coordinating Centre Tutors (CCTs) conduct in-service teacher trainings and provide support and supervision of teachers through classroom monitoring visits that involve providing feedback and guidance on teaching quality and best practices. The TDMS often works with a cascade or "train-the-trainer" approach to training, intending that trainers pass on skills and competences to CCTs, who then directly train teachers.

The limited resources, low levels of teacher training and support, overall rates of poverty, and low levels of adult literacy throughout the country, coincides with severe educational problems in Uganda. After three years of completed primary education 41% of children still cannot read a single word, and by Grade 7 this number is still 10% (Uwezo 2016). To graduate from primary school, students must take the Primary Leaving Exam (PLE). Only 53% of students passed the PLE in 2017.

2.2 School Closures during COVID-19

Uganda had one of the longest school closures i the world during COVID-19 with 66 weeks of full school closure and 23 weeks of partial closure. Schools where fully closed between March 2020 and October 2020, then reopened for a couple of months until end of December 2020. Full closure again from January to March 2021 and re-opened between March and June 2021. From June 2021 to January 2022 schools where fully closed again. Finally, schools started to fully re-open on January 10^{th} 2022 (UNESCO 2023). This means that many children were out of school for almost *two* years. When returning to school the official policy was that children should be automatically promoted from the grade they where in when schools closed in 2020 to the next respective grade. So for example, if a student was in P4 in 2020 they should be in P5 in 2022. This means that "on track" progression was delayed by one year after COVID-19.

2.3 Northern Uganda Literacy Program

The Northern Uganda Literacy Project (NULP) was developed by Mango Tree Educational Enterprises Uganda, a private, locally-owned educational tools company that developed an instructional methodology to improve early-grade literacy, with a focus on mother-tongue-first instruction. In 2009, Mango Tree began developing their literacy program for one language group, Leblango. The NULP was piloted and refined from 2009 to 2012 to determine what was pedagogically and logistically successful. The project was based in the Lango sub-Region, where the vast majority of the population speaks Leblango.

Although the national policy is to instruct early primary students in their local language, in practice, with 41 different languages from three different language families, poorly developed orthographies, lack of teaching and reading materials, and deficient teacher training, English is still heavily used in primary schools across Uganda. Rote memorization of how to read basic words (in English) aloud is a common technique in early primary classrooms. The NULP model involved a revised curriculum for grades 1-3 that explicitly instructed teachers how to teach in the local language and avoided using any written English text during grade one. Moreover, NULP introduced content more slowly than the standard curriculum; approximately half of the letter/sounds of the alphabet to be taught in grade one, with the remaining taught in grade two.

The NULP provided extensive training and support for teachers in participating classrooms using expert trainers, detailed facilitator's guides, and instructional videos. The program provided teachers with a five-day residential (off-site) training in Leblango orthography and literacy prior to the beginning of the school year. Teachers also underwent two additional intensive, residential trainings on literacy methods during the breaks between the three school terms. In addition to the residential trainings, there were also six in-service training workshops on Saturdays throughout the school year. The training sessions were complemented by support supervision visits that provided teachers feedback about their teaching. These were conducted three times each term by Mango Tree staff members. Teachers also received two support supervision visits from CCTs each term. The CCTs were trained in providing the same sort of feedback on teaching performance as the Mango Tree staff.

In addition to teacher training, NULP provided each classroom with a set of materials tailored for their instructional model including primers and readers. Teachers were provided with scripted lesson guides for each literacy lesson. Grade one classrooms were provided with slates that allowed each student to practice writing individually using pieces of chalk. Each classroom was provided with a wall clock to help teachers keep track of the time during a lesson.

Finally, NULP helped to support school parent meetings once per term and actively

engaged with communities more broadly to increase communities' knowledge of, and appreciation for, their written local language. This included a radio program that conducted literacy and local language promotion. The radio program and local language promotion took place across the region and thus we cannot evaluate their effects in this paper.

3 Evaluation and Data

3.1 NULP Evaluation

To assess the impact of the NULP on student learning, we conducted a multi-year, randomized evaluation of the program. Schools were sampled for the study in two phases: an initial pilot RCT in 2013, and a larger RCT in 2014 which continued in 2015 and 2016. In 2013, 38 eligible schools were selected to be part of the RCT. To be eligible, schools had to meet a set of criteria established by Mango Tree, the most important being that each school needed to have exactly two P1 classrooms and teachers. In 2014 the program was expanded to 90 additional schools for a total of 128 schools. The eligibility criteria for these new 90 schools were slightly different, and less stringent.[3]

Of the 128 schools, the evaluation assigned each to one of three study arms: 1) Full-cost NULP, 2) Reduced-cost NULP, and 3) Control. In the Full-cost arm, schools received the original NULP as designed by and delivered by Mango Tree and its staff. In the Reduced-cost arm, some of the materials (slates and chalk) were eliminated, training was conducted through a cascade model led by government employees (Ministry of Education staff) rather than Mango Tree staff, and teacher received fewer support visits, again from government employees.[4] Schools in the Control group did not receive the literacy program. To randomize treatment assignments, schools were grouped into stratification cells of three schools each. Each stratification cell had its three schools randomly assigned to the three different study arms via a public lottery.

This study focuses on the cohort of students who entered the 128 study schools as first-graders in 2014.[5] The NULP was introduced to first-grade classrooms and teachers in treatment schools in 2014. In 2015, the NULP was directed at second-grade classrooms and teachers, and in 2016 the program was directed at third-grade classrooms and teachers.[6]

[3] Criteria in 2014 include: having desks and blackboards in grade 1-3 classrooms and having a student-to-teacher ratio of no more than 150 students during the 2013 school year in grades P1 to P3.

[4] Kerwin and Thornton (2021) discuss and quantify the differences between the full- and reduced-cost program versions, and present the results of the 2013 pilot RCT in 38 schools.

[5] The treatment schools from the original 38 schools received the program in P1 in 2013, and then repeated the program P1 in 2014. For the other 90 schools, the program was first introduced in 2014.

[6] In 2017, Mango Tree piloted a teacher mentor program with fourth-grade teachers in the reduced-cost

Classrooms were allowed to keep all of the Mango Tree educational materials (such as slates, primers, and readers) in the years after they initially received the program, but teachers no longer received additional training or support visits. This means that children starting Grade 1 in 2014 who progressed an additional grade each year, were treated for three years. To assess learning gains, 100 P1 students were sampled at random from each of the 128 schools.[7] This cohort was tracked into P2 and P3 in 2015 and 2016, respectively.

3.2 Tracking Students to Measure Lasting Impacts

3.2.1 Sample and Data Collection

To assess the longer-term effects of the NULP we collected data between March 2021 and June 2022 for a targeted subset of the original cohort of students who were in P1 in 2014. The sampling approach was designed to minimize attrition while maintaining a valid sampling frame balanced across study arms. To achieve this, we restricted the sample in the following ways. First, we selected only students who were sampled into the study at the beginning of the school year (baseline) in 2014, excluding those who were sampled at the end of that school year. Second, we included only students who were randomly sampled for a parent survey conducted in 2015, which gives us additional information critical for tracking the students. Third, we excluded all students from stratification cells that were in Lira District.[8] After applying these criteria the target follow-up sample consisted of 3,108 students in 104 schools, which is slightly under half of the original sampled children in 2014 and four-fifths of the original 128 schools.

Prior to data collection, we held two school/community meetings at the end of 2020 to update information on the current location of the sampled children and mobilize parents for data collection. These efforts to locate the children resulted in only 10% of the sampled children categorized as "Not found" at this stage. Of the 2,788 students who could be located, 64% were still in the expected school/community and 26% were in a nearby school/community.

and full-cost schools to provide support; no materials or pedagogical training or support were delivered. This intervention was much less intensive than the earlier years.

[7] The sampling procedure differed slightly between the original 38 schools and the 90 schools added in 2014. In the 38 schools that participated in 2013, an initial sample of 40 grade one pupils was drawn at the 2014 baseline, and then 60 students were added at the 2014 endline following the same sampling procedure as at baseline. In the 90 new schools, 80 students were selected at baseline with an additional 20 added at endline. The difference was due to the organizational difficulty of testing large numbers of students at baseline or endline at each school, since the study also collected data on the second-grade students from the original 38 schools that had been exposed to the program in 2013.

[8] This last exclusion was imposed because families from the city tend to move more, and thus tracking would be more challenging and costly in that area. Since the treatment assignment was randomized within each stratification cell, excluding these cells does not internally bias our estimates.

According to official government policy, our cohort of students should have been in grade 7 in 2021, delayed one year relative to their initial trajectory due to the pandemic-related school closures. Grade repetition would result in students enrolled in far lower grades. We targeted students in grades 7 and lower for these reasons. Due to the restrictions during the COVID-19 pandemic, the data collection process was divided into three phases; Phase 1: March 6rd to March 18th 2021 (grades 6 and 7)[9], Phase 2: November 6th to December 11th 2021 (grades 5 to 1), and Phase 3: May 26th to June 14th 2022 (students that could not be found in the two previous phases). Data collection resulted in a total of 2,314 children; the attrition rate out of the overall targeted sample of 3,108 was 25.5%.

5 March 2021 and ended on 14 Jun 2022, with a median date of 16 Nov 2021 and a mean date of 7 Oct 2021. So I think we should say our followup was 8-9 years post-baseline, and 5-6 years after the treatment ended.

3.2.2 Data

We collected two types of data: learning assessments in Leblango, English, and math, and a student survey containing information on grade progression, school dropout, progression to secondary school, working outside of the home, and sexual behavior.

Reading skills in Leblango and English were assessed through the Early Grade Reading Assessment (EGRA), which is an internationally standardized exam designed to assess early literacy skills such as recognizing letters, reading simple words and understanding sentences and paragraphs. We use an adaptation of the EGRA to Leblango, which covers two components of literacy skills: oral reading fluency (ORF), and reading comprehension (RC). Math skills are measured through the Early Grade Math Assessment (EGMA), which covers three components of basic math skills; Addition, Subtraction and Numerical problems. In addition to the learning assessments collected in 2021 we also use the assessments collected at baseline in 2014.

To measure overall performance across all components of the assessments, we construct a principal components score index for each of the three subjects in the following way. For each assessment (English and Leblango EGRA, and EGMA), we estimate the weights of each skill in the first principal component for the control-group students in the 2014 baseline evaluation. Then we use those weights to predict this principal component among treated students in the baseline evaluations, as well as all students in the follow-up. We then standardize each index by subtracting the control mean and dividing by the control-group standard deviation in each year (2014 and 2021).

[9] As part of the gradual re-opening of school, grades 7 and 6 were allowed back to school earlier than grades 5 and down.

In addition to the standardized index, we rescale the learning gains into Equivalent Years of Schooling (EYS) as describe by Evans and Yuan (2019) which communicates how many additional "business-as-usual" years of schooling are achieved for the given learning outcomes. To do so, we compute the annual gain in SDs for the control group by subtracting the baseline score from the endline score and dividing with the average number of grades advanced by control students. The grades advanced by students is, at most, the number of years elapsed between the baseline and endline evaluations. For Leblango and Math this is maximum three years (from beginning of 2014 to end of 2016 for the immediate effects or to the end of 2021 for the five-year follow-up) and for English this is maximum one year (from end of 2015 to the end of 2016 for the immediate effects or to the end of 2021 for the five-year follow-up). Then we calculate EYS by dividing the regression estimate β_1 or β_2 from Equation (1) with the annual gain of the control group

School attendance is measured as a zero-one indicator where one indicates that the student ever attended school in a given year, as measured by the student survey responses. School dropout is also measured as a zero-one indicator where a one indicates that the student ever dropped out of school in a given year. Note that students often drop out of school and return within a single year, so these two indicators are not mutually exclusive. Grade progression measures which grade the child attends in a given year. Attending secondary school is measured as a zero-one indicator where a one indicates that the child attended secondary school in a given year. Working outside the home is measured as a zero-one indicator where one indicates if the child worked outside the home in a given year. All of these measures are retrospective and asked of the child in the 2021 survey, for all years from 2014 through 2021. Sexual experience is measured through three variables: ever had sex (where one indicates yes), age at sexual debut, and sexual debut before age 13 (where one indicates yes).

4 Empirical Strategy

Following our analysis plan, we obtain experimental impact estimates of the treatment effects for each of our outcomes y_{ij} via the following parametric linear model estimated by ordinary least squares:

$$y_{ij} = \beta_0 + \beta_1 \mathrm{FC}_j + \beta_2 \mathrm{RC}_j + \mathbf{Z'}_j\tau + \mathbf{X'}_i\gamma + \varepsilon_{ij} \tag{1}$$

where i indexes students, who attended school j in grade one in 2014. FC_j and RC_j are indicators for a school being randomly assigned to the Full-cost or Reduced-cost NULP program, respectively. Z_j is a vector of indicators for the stratification cells used in the

random assignment of schools to study arms. X_i is a vector of control variables and include an indicator for being male, age at baseline (inputted as categorical indicators for each age)[10], and baseline test score indices for Leblango, math, and oral English.[11] We conduct inference on our estimates via randomization inference. Specifically, we randomly permute the study arm assignments of each school within the stratification cells used in the original random assignment.[12]

We consider three null hypotheses for each outcome we study:

$$H_0^1 : \beta_1 = 0 \tag{2}$$

$$H_0^2 : \beta_2 = 0 \tag{3}$$

$$H_0^3 : \beta_1 = \beta_2 \tag{4}$$

In our pre-analysis plan we established three broad sets of outcomes we would study: 1) learning outcomes (Leblango EGRA, English EGRA, and EGMA – all measured in SDs), 2) downstream academic outcomes (attended school in 2021, attended secondary school in 2021), and 3) exploratory life outcomes (ever had sex, first had sex at age 13 or below, worked outside of the home in 2021, and worked outside of the home in a non-agricultural sector in 2021). In Section 5, we present the results for these outcomes, but we deviate from the plan in two ways. First, we adjust the way in which we construct EGRA and EGMA scores by only including the test components which were measured in the 2014 baseline and 2021 endline assessments. Second, In addition to the standardized index, we rescale the learning gains into EYS in order to give a better sense of the magnitude of the effect. Third, we changed the set of exploratory outcomes during the analysis process, as we discovered important results that were not considered in the analysis plan. The results that exactly follow the analysis plan are in Appendix A.3 and do not differ substantively from those in this paper.

In Appendix A.3 we also take account of multiple hypothesis testing using the Benjamini, Krieger, and Yekutieli (2006) method to compute sharpened q-values that control the false discovery rate (FDR). Following Derksen et al. (2023), we use the Anderson (2008) implementation of their approach, which computes the lowest value of the sharpened q-value for

[10] We bottom-code age at 5 and top-code it at 9.

[11] The baseline exams were conducted at the beginning of the 2014 school year. For students with missing values of the baseline exam score, we replace the missing values with zero and include a separate indicator variable for the baseline exam score being zero.

[12] We implement this in Stata via the *ritest* command, following the approach in Kerwin and Thornton (2021).

which we can reject the null, so that our q-values can be interpreted in the same way that conventional p-values are[13]. We adjust for multiple testing separately by domain.

4.1 Attrition

One important issue to address when evaluating the impact of a program over a longer time horizon is attrition. As mentioned in Section 3, we were able to track 74.5% of our sample in 2021. Attrition is a concern for internal validity if it is correlated with treatment assignment, and also a potential concern for external validity if correlated with student, household, or school characteristics.[14] In Table 1, we show the correlation of attrition by treatment status by regressing an indicator of attrition on the two treatment indicators using the same specification as Equation (1), specified above. For completeness, we examine the probability of attrition from either taking the EGRA and EGMA tests (Table 1, column 1) and from completing the student survey (Table 1, column 2). Although attrition rates are between 25.5% and 28.1% in the control group, there are no statistically significant differences in attrition across the treatment arms.

In Table 2, we include baseline co-variates to the regressions to understand the overall predictors of attrition and whether the type of attritors vary across treatment groups. In addition to including the two treatment group We also use a more flexible approach by allowing for interaction between the treatment indicator and all the baseline covariates and joint testing for the statistical significance of all the interactions of each treatment arm. The results for the students who attritted from testing are presented in Table 2. The results in Table 2 stems from one single estimation where column (1) presents the level effects and columns (2) and (3) present the interaction between the level and being in the reduced-cost and full-cost program, respectively. Table 2 shows that none of the sets of variables (baseline covariates in levels, the interactions with the reduced-cost treatment, and the interactions with the full-cost treatment) pass a joint significance test.

These results suggest that even though we were not able to follow about a fourth of the sample of students, the attrited children do not appear to be different than the rest in terms of our set of baseline covariates. Furthermore, neither of the treatments affects the probability of attriting from the sample. This implies that the risk of attrition causing bias in our estimates is low.

[13] We thank Olivier Sterck for sharing the code used to estimate these q-values.

[14] For example, if attrition is correlated with academic ability – ie. poor performing students are most likely to attrit, and if treatment effects vary by this characteristic, our effects could over- or under-state the true effect on the entire population of students.

Table 1
Test of Differential Attrition

	(1) Attrited from testing	(2) Attrited from testing and survey
Reduced-Cost	-0.005	-0.017
S.E.	(0.024)	(0.026)
R.I. p-value	[0.887]	[0.619]
Full-Cost	0.007	0.003
S.E.	(0.023)	(0.024)
R.I. p-value	[0.799]	[0.913]
Difference Between Treatments	0.012	0.020
S.E.	(0.025)	(0.026)
R.I. p-value	[0.739]	[0.587]
Control Mean	0.255	0.281
Control SD	0.436	0.450
N	3108	3108

Notes: The outcome is a zero-one variable indicating if the student has attritted. In column (1) attrition is defined as not being tested and in column (2) attrition is defined as not being tested nor answering the student survey. Covariates are measured as of the baseline exams in 2014. All regressions control for stratification cell indicators and baseline values of Leblango EGRA Score, Oral English assesment, EGMA, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in brackets.

Table 2
Test of Differential Attrition with Interactions

	(1) Levels	(2) Reduced-cost Interaction	(3) Full-cost Interaction
Reduced-cost	-0.084		
	(0.137)		
Full-cost	0.069		
	(0.114)		
English EGRA score	0.028**	-0.043**	-0.005
	(0.013)	(0.020)	(0.019)
English EGRA score missing	-0.199*	-0.009	0.084
	(0.116)	(0.186)	(0.199)
Leblango EGRA score	-0.016	0.003	0.021
	(0.022)	(0.030)	(0.023)
Leblango EGRA score missing	0.074	0.044	-0.137
	-0.102	(0.143)	(0.152)
Math score	0.000	0.006	-0.023
	(0.019)	(0.025)	(0.025)
Math score missing	0.015	-0.023	0.015
	(0.030)	(0.049)	(0.045)
Male	0.008	0.038	0.04
	(0.031)	(0.038)	(0.037)
Age 5 or lower	-0.074	0.032	-0.055
	(0.098)	(0.136)	(0.141)
Age 6	-0.115	0.072	-0.094
	(0.072)	(0.131)	(0.108)
Age 7	-0.067	0.054	-0.103
	(0.070)	(0.127)	(0.111)
Age 8	-0.083	0.000	-0.072
	(0.075)	(0.128)	(0.113)
Age 9	-0.063	0.176	-0.108
	(0.090)	(0.137)	(0.121)
Joint p-value Interactions	[0.896]	[0.275]	[0.896]

Notes: All results stem from one single estimation, where column (1) presents the level effects and columns (2) and (3) present the interaction between the level and being in the reduced-cost and full-cost program, respectively. The outcome is a zero-one variable indicating if the student has attritted from testing (i.e. corresponding to the sample in Table 1 column (1)). N is 3108 students of which 793 attritted from testing. Covariates are measured as of the baseline exams in 2014. All regressions control for stratification cell indicators and baseline values of Leblango EGRA Score, Oral English assesment, EGMA, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in brackets.

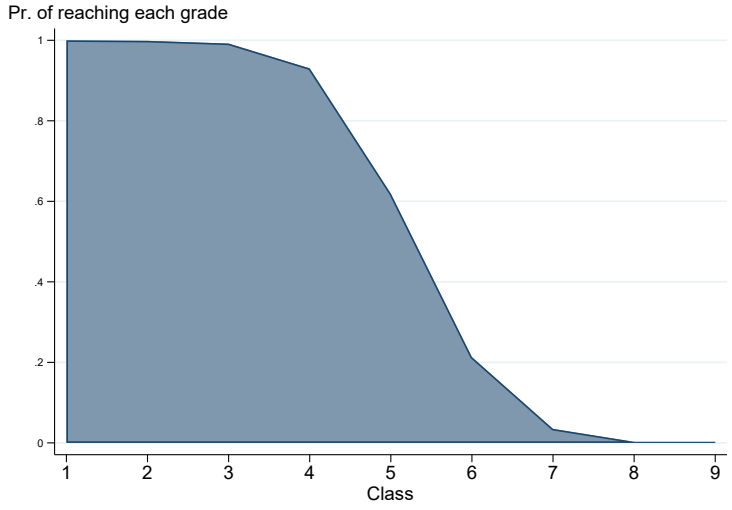
5 Results

We present three sets of results. First, we examine how educational outcomes progress from 2014 to 2021 among control group students, and the impact of the COVID-19 pandemic. We do this by following the enrollment and grade progression of the students in the control group. Second, we estimate the treatment effects of the NULP measured in 2016–immediately after the program ended. Finally, we present the longer term treatment effects measured in 2021–five years after the program ended.

5.1 Longitudinal Results amongst the Control Group

In this section, we follow the cohort of control students that were in grade 1 in 2014 and show their school progress up until 2021. Figure 1 shows the probability of reaching each grade from grade one in primary school to standard 1 in secondary school (grade 8 in Figure 1). Figure 1 shows that almost every student that started in 2014 have by 2021 completed grade three and many have also completed grade four. From grade four the probability of completing higher grades sharply declines and virtually none have made it to secondary school. Even with one year delay due to the COVID-19 school closures we would have expected most children to have reached at least grade 6 as they should have done so in 2019 before the onset of COVID-19, yet only 20% have.

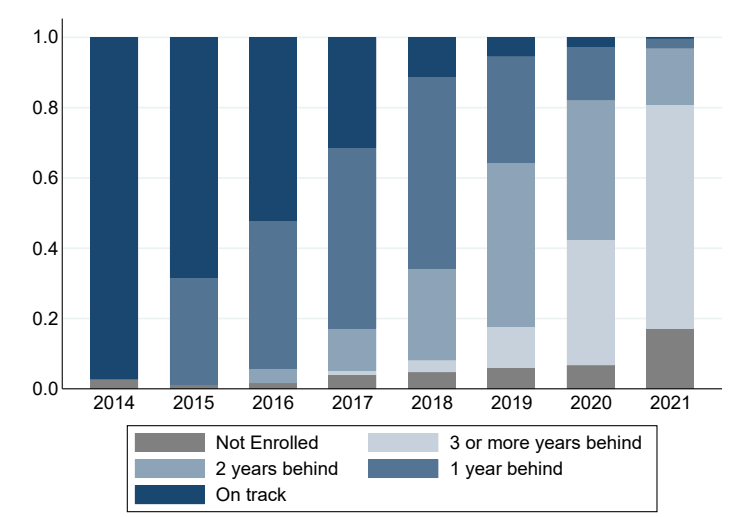
Figure 1
Probability of Control Students Reaching each Grade by 2021



Notes: This figure shows the probability that students in the control group reach each grade by 2021. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

To better understand the lack of progression showed in Figure 1 we in Figure 2 show how many years students are behind each year. Figure 2 shows that the lack of progression starts early and that 30% have repeated Grade 1. In 2016 most children have completed grade 1 (two years later than expected). However, the fraction of children who are "on track" steadily declines from 2014 onward and in 2019 less than 10% are "on track". In 2019 the majority of students are two years behind their expected grade level.

Figure 2
Grade Progress by Control Students each Year



Notes: This figure shows the degree of repetition in the control group was each year, as remembered by the student in 2021.

Table 3 presents this pattern of grade repetition in more detail and shows that up until 2016 the majority of children are "on track". However, from 2017 we see that more than half of the children are behind their expected grade level and in 2019 only 5% are "on track".[15]

Table 3
Grade Attended by Control Students each Year

Grade Attended	Year							
	2014	2015	2016	2017	2018	2019	2020	2021
Not enrolled	0.03	0.01	0.02	0.04	0.04	0.06	0.06	0.17
1	0.97	0.31	0.04	0.01	0.00	0.00	0.00	0.00
2	0.00	0.68	0.42	0.12	0.03	0.01	0.00	0.00
3	0.00	0.00	0.52	0.52	0.27	0.11	0.05	0.04
4	0.00	0.00	0.00	0.31	0.55	0.47	0.31	0.25
5	0.00	0.00	0.00	0.00	0.11	0.30	0.41	0.36
6	0.00	0.00	0.00	0.00	0.00	0.05	0.15	0.16
7	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

Overall, we see that grade repetition is the norm — 93% of all students report to have repeated at least one grade (see Table A6). Very few repeated Grade 1 (in 2014) but otherwise grade repetition is between 20 and 50 percent. Moreover, we see the most students report to have repeated one to three grades in total, only 7% have never repeated a grade, and 11% have repeated four grades (see Table A8). Looking at which grades are most frequently repeated we in Table A7 see that the probability of repeating is increasing in grades up until grade four with a grade repetition of 72%. The reason for grade repetition to decline after Grade 4 is not that children are not repeating but rather that our cohort of kids have not reached the higher grades yet.

Exploring reasons for this massive grade repetition we grouped open-ended answers into categories each year. Table 4 presents the three main reasons for repeating a grade each year.

In 2014 where our sample of kids started in Grade 1 the most frequent reason for repeating was that the child was considered too young to be promoted to the next grade. This reason was also present in 2015 but then wasn't present in later years when kids got older. Between

[15] This pattern doesn't change much if we divide the sample by gender (see Table A1 and Table A2) nor initial academic performance (see Table A3 and Table A4)

Table 4
Top Three Main Reasons for Grade Repetition

2014	2015	2016	2017	2018	2019	2020	2021
Age-related Challenges	Academic Challenges	Academic Challenges	Financial Constraints	Academic Challenges	Financial Constraints	Financial Constraints	COVID-19 Impact
Academic Challenges	Financial Constraints	Financial Constraints	Academic Challenges	Financial Constraints	Academic Challenges	Academic Challenges	Financial Constraints
Health Issues	Age-related Challenges	Health Issues	Health Issues	Advised to repeat	Health Issues	COVID-19 Impact	Academic Challenges

Notes: The first row shows the most frequent reason for children repeating a grade in each year. The second row shows the second most frequent reason and the third row the third most frequent.

2015 and 2019 academic challenges and financial constraints was the two main reasons for children to repeat. Both reasons are related to the practice of promotional exams between grades. Even though promotional exams are officially illegal it is a widespread practice. Sitting for the promotional exam requires a fee so if parents are not able to pay the child will repeat that grade. Relatedly, if a child is unlikely to pass the test the teacher can advise not to sit for the promotional exam and thus repeat instead. In 2020 and 2021 the onset COVID-19 becomes a substantial reason for repeating.

In this section we have shown that in the status quo children are lacking behind their expected grade levels and the majority ends up being two grade levels behind. In the next two sections we present results of the impact of the NULP on learning, grade progression and life-outcomes.

5.2 Immediate Impacts of the NULP on Learning

The NULP intervention was implemented over three years (grades 1 to 3) running from 2014 to 2016. Table 5 presents the immediate impact of both the full-cost and reduced-cost versions of the NULP on student achievement. Columns (1) to (4) present the treatment effects on Leblango (local-language) and English reading, respectively, and columns (5) and (6) presents the impact on math. Students in the reduced-cost version did 0.7 SDs better in Leblango reading compared to the control group students and 0.3 SDs better in English. There is no statistically significant difference in math achievement.

Moving to the full-cost version students scored 1.2 SDs higher in Leblango reading compared to the students in the control schools and 0.5 SDs higher in English reading. As for the reduced-cost, there was no difference in math performance between the full-cost and the control. The difference between the two versions of the NULP is also statistically significant.

The magnitude of these gains may be easier to appreciate in terms of equivalent years of

Table 5
Immediate Impacts on Leblango, English and Math Test Scores

	Leblango		English		Math	
	(1)	(2)	(3)	(4)	(5)	(6)
	SD	EYS	SD	EYS	SD	EYS
Reduced-cost	0.707***	4.234***	0.277***	0.780***	-0.081	-0.132
S.E.	(0.091)	(0.545)	(0.078)	(0.219)	(0.056)	(0.092)
R.I. p-value	[0.000]	[0.000]	[0.006]	[0.006]	[0.227]	[0.227]
Full-cost	1.227***	7.346***	0.515***	1.450***	-0.058	-0.095
S.E.	(0.096)	(0.572)	(0.085)	(0.240)	(0.059)	(0.097)
R.I. p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.471]	[0.471]
Difference Between Treatments	0.520***	3.112***	0.238***	0.671***	0.023	0.038
S.E.	(0.107)	(0.643)	(0.070)	(0.197)	(0.057)	(0.093)
R.I. p-value	[0.000]	[0.000]	[0.003]	[0.003]	[0.761]	[0.761]
Control Annual Gain (in SD)	0.167		0.355		0.611	
Control Mean	0.000	2.993	0.000	0.995	-0.000	2.897
Control SD	1.000	0.097	1.000	0.081	1.000	0.397
Full-cost Mean		10.194		2.407		2.939
N	6114	6114	6109	6109	5954	5954

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Sample includes all students present in the 2016 follow-up.

schooling. Columns (2), (4) and (6) present these results for Leblango, English and math, respectively. *completed grades* is 2.9 for Leblango and math. For the English EGRA, control-group students had progressed 0.9 grades between tests on average (from end of 2015 to end of 2016). Column (2) shows that for Leblango, the students who received the reduced-cost treatment gained equivalent to 4.2 additional years of learning, and students in the full-cost treatment arm gained equivalent to 7.4 additional years of learning. This means that students in the reduced-cost group learned more than twice as much and the students in the full-cost group learned more than three times as much compared to the control group students. The difference between the full-cost and reduced-cost versions of the program is high enough that we can reject the null hypothesis that the gains are equal. For English (Column (4)) we see the same pattern although smaller magnitudes. As English starts later the students only had one year of English reading by the end of 2016. For students in the reduced cost version they gained equivalent to 0.78 additional years of schooling and the students in the full-cost treatment arm gained equivalent to 1.45 additional years of schooling. This means that the stundents in the reduced cost group gained almost twice as much learning while til students in the full-cost gained more that twice as much learning compared to the control group.

Another way of presenting the magnitude of the impacts is by oral reading fluency which

is a common international metric. Oral reading fluency is measured in words per minutes indicating whether the child is able to read at a pace where it is possible to exact meaning from the text.

Table 6

Immediate Impacts on Leblango Words Read per Minute

	(1) WPM	(2) WPM SD	(3) WPM>0	(4) WPM>20	(5) WPM>45
Reduced-cost	7.428***	0.705***	0.269***	0.151***	0.055***
S.E.	(0.959)	(0.091)	(0.027)	(0.022)	(0.010)
R.I. p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Full-cost	12.933***	1.227***	0.379***	0.272***	0.114***
S.E.	(1.006)	(0.095)	(0.029)	(0.023)	(0.012)
R.I. p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Difference Between Treatments	5.504***	0.522***	0.110***	0.121***	0.059***
S.E.	(1.134)	(0.108)	(0.026)	(0.025)	(0.013)
R.I. p-value	[0.000]	[0.000]	[0.000]	[0.001]	[0.003]
Control Mean	5.245	-0.000	0.594	0.090	0.008
Control SD	10.537	1.000	0.491	0.287	0.090
N	6114	6114	6114	6114	6114

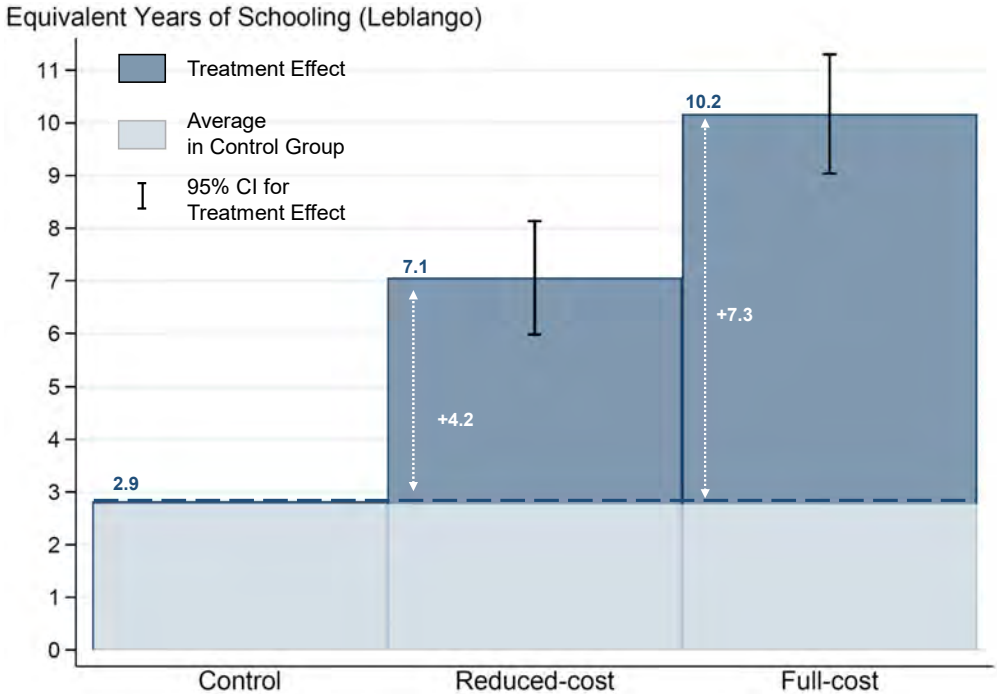
Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference *p*-values, clustered by school (128) and stratified by stratification cell (36), in square brackets []. Sample includes all students present in the 2016 follow-up.

Table 6 presents the impact of the NULP on different cut-offs of oral reading fluency for Leblango. Column (1) presents the average impact and shows that children in the full-cost treatment group could read around 13 words more per minute compared to the control group for the reduced-cost treatment the increase was 7 words per minute. On average the control group students could read only 5 words per minute which is well below any standard for being considered able to read. This low average could be due to a large fraction of non-reading children. Thus in the next columns we focus on different thresholds of oral reading fluency. In Column (3) we see that 60% of the children can read at least one word (making 40% of the children non-readers). The full-cost program reduced the group of non-readers to almost zero while in the reduced-cost group 10% were still non-readers. In Column (4) we see that only 9% of the children in the control group can read more than 20 words per minute (which is considered a minimum fluency threshold in other African languages (Ardington et al. 2021)). This percent increase by 27 percentage points in the full-cost group and 15 percentage points in the reduced-cost group. Finally in Column (5) we consider a threshold (45 wpm) which is the level children should be able to read at at the end of third grade (Ardington et al. 2021). However, only 1% of students in the control group were able to read

at this level. For the full-cost group this increases to 12% and for the reduced-cost 6.5%. Figure A1 shows that the impact of NULP on oral reading fluency is highest for the lowest performing children and turns to zero for the highest performing children.

Figure 3

Immediate Treatment Effects on Leblango in Equivalent Years of Schooling



Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Whiskers represent heteroskedasticity-robust 95% confidence intervals, clustered by school. Detailed results are presented in Table 5 column (2).

5.3 Impacts of the NULP at Five Years Follow-up

Table 7 presents the treatment effects of the full-cost and reduced-cost versions of the NULP on student test scores eight years after the program started and five years after it ended. We present these estimates in two forms. Columns (1), (3), and (5) measure the effects in SDs of the EGRA (Leblango and English) and EGMA (math) assessments.

We find a large long-run impact of the NULP on student learning in Leblango and English, but not in math. Five years after the intervention ended, students who received the reduced-cost version of the program scored 0.38 and 0.25 SDs higher than the control students on the Leblango and English EGRA, respectively. More impressively, the full-cost version of the program had sustained impacts on student learning of 0.71 and 0.55 SDs, respectively. The math estimates show a small and statistically insignificant negative impact of -0.04 SDs for

Table 7
Treatment Effects on Leblango, English and Math Test Scores, (Five Year Follow-up)

	Leblango		English		Math	
	(1)	(2)	(3)	(4)	(5)	(6)
	SD	EYS	SD	EYS	SD	EYS
Reduced-cost	0.377***	2.341***	0.249***	0.724***	-0.036	-0.110
S.E.	(0.148)	(0.918)	(0.101)	(0.293)	(0.045)	(0.136)
R.I. p-value	[0.000]	[0.000]	[0.001]	[0.001]	[0.527]	[0.527]
Full-cost	0.712***	4.422***	0.547***	1.590***	-0.000	-0.001
S.E.	(0.200)	(1.241)	(0.141)	(0.408)	(0.041)	(0.124)
R.I. p-value	[0.000]	[0.000]	[0.001]	[0.001]	[0.996]	[0.996]
Difference Between Treatments	0.335	2.081	0.298	0.867	0.036	0.109
S.E.	(0.213)	(1.323)	(0.153)	(0.445)	(0.039)	(0.119)
R.I. p-value	[0.237]	[0.237]	[0.117]	[0.117]	[0.470]	[0.470]
Control Annual Gain (in SD)	0.161		0.344		0.332	
Control Mean	0.000	4.796	-0.000	3.117	0.000	4.796
Control SD	1.000	6.211	1.000	2.907	1.000	3.012
Full-cost Mean		9.137		4.790		4.860
N	2315	2315	2315	2315	2314	2314

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets [].

the reduced-cost version and a point estimate of exactly zero (to three decimal places) for the full-cost version. We can reject impacts on math larger than 0.08 SDs, which is smaller than the average effect size of 0.1 SDs for education RCTs in the developing world (Evans and Yuan 2022).

In terms of equivalent years of schooling we see that even though eight years had elapsed from the beginning of intervention (from 2014 to 2021), the number of *completed grades* is far less than eight. The average student in the control group had only completed almost five grades since the beginning of 2014 (Table 7 Columns (2) and (6)). For the English EGRA, control-group students had progressed around three grades between tests on average, despite the fact that six years have elapsed (from end of 2015 to end of 2021). Table 7 Column (2) shows that students in the reduced-cost group are now equivalent to 2.3 years ahead and students in the full-cost group equivalent to 4.4 years ahead. This means that the control group students have to some degree catched up compared to the results in 2016. However, despite this the full-cost students have still gained more than twice as much learning compared to the control group. Column (4) presents the results for English and shows that the immediate gains in 2016 are preserved five years later. Although the estimates for the full-cost version of the program are higher than those for the reduced-cost version in all cases, we cannot reject the null hypothesis that the gains are equal.

Table 8 presents the results on oral reading fluency. Column (1) shows that five years after the program ended the control group students were able to read 14 words per minute (compared to 5 words five years prior) this means that on average students in the control group increased their reading speed with approximately two words per year. Column (3) shows that the group of non-readers haven't changed for the control group between 2016 and 2021. The group of non-readers are reduced for both the full-cost and reduced-cost although at a lesser extent than in 2016 [16]. In columns (4) and (5) we see that the fraction of students in the countrol group that can read more than 20 wpm and more than 45 words per minute has increased since 2016, but also that students in both the full-cost and reduced-cost are still ahead of students in the control group. Similar to earlier we find that the impact on oral reading fluency decreases with the achievement level of the student (see Figure A2).

Table 8

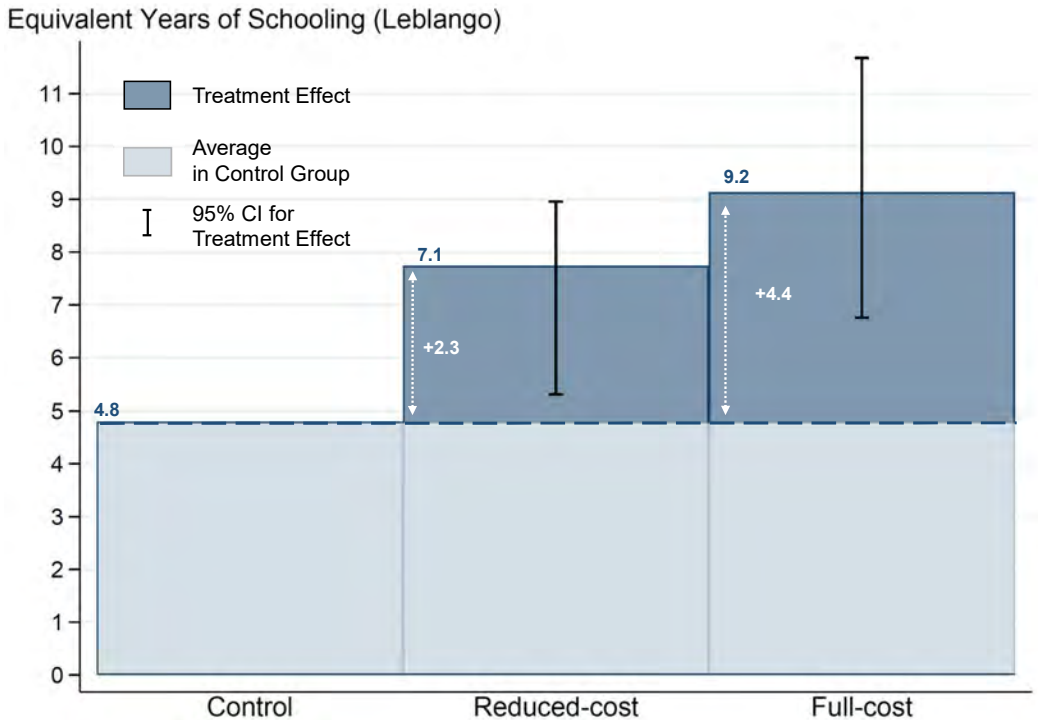
Treatment Effects on Leblango Words Read per Minute (Five Year Follow-up)

	(1) WPM	(2) WPM SD	(3) WPM>0	(4) WPM>20	(5) WPM>45
Reduced-cost	7.403***	0.385***	0.094***	0.096***	0.068***
S.E.	(3.032)	(0.158)	(0.027)	(0.024)	(0.019)
R.I. p-value	[0.000]	[0.000]	[0.007]	[0.000]	[0.006]
Full-cost	14.153***	0.737***	0.126***	0.138***	0.111***
S.E.	(4.102)	(0.214)	(0.029)	(0.030)	(0.021)
R.I. p-value	[0.002]	[0.002]	[0.006]	[0.002]	[0.000]
Difference Between Treatments	6.750	0.351	0.032	0.042	0.043
S.E.	(4.379)	(0.228)	(0.029)	(0.029)	(0.023)
R.I. p-value	[0.227]	[0.227]	[0.363]	[0.263]	[0.151]
Control Mean	14.013	0.000	0.551	0.308	0.094
Control SD	19.207	1.000	0.498	0.462	0.292
N	2315	2315	2315	2315	2315

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference *p*-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Sample includes all students present in the 2016 follow-up.

[16] Note that the sample has changed so the could also just be an artifact of the change in sample

Figure 4
Treatment Effects on Leblango in Equivalent Years of Schooling
(Five Year Follow-up)



Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Whiskers represent heteroskedasticity-robust 95% confidence intervals, clustered by school. Detailed results are presented in Table 7.

Table 9 shows the effect of the NULP on self-reported school attendance in 2019, 2020, and 2021. Almost all control-group students report attending school during 2019 and 2020 (95% and 94%, respectively), but attendance dropped to 84% in 2021 due to the Covid-19 pandemic. Reported attendance in 2021 is surprisingly high considering that all government schools in Uganda were closed for the vast majority of that year. This leads us to believe that students reported attending even if they actually only went to their schools for a small number of days those years.

We find no statistically significant effect of the NULP on attending school in any of the three years. The point estimates for the reduced-cost version of the program were 0.7, 1.3 and 1.6 percentage points for 2019, 2020 and 2021 respectively, while the full-cost point estimates were -1.3, -0.5 and 0.1 percentage points for the same years. In these analyses we have 80% power to detect effects of 3.1 percentage points in 2019 and 2020 and 6.3 percentage points in 2021. We thus cannot rule out the possibility of small effects on attendance in 2021.

We originally planned to estimate the effects of the NULP on the probability of attending secondary school. However, we found only four students who managed to progress to that

Table 9
Treatment Effects on School Attendance

	(1) 2019	(2) 2020	(3) 2021
Reduced-cost	0.007	0.013	0.016
S.E.	(0.010)	(0.011)	(0.025)
R.I. p-value	[0.599]	[0.367]	[0.575]
Full-cost	-0.013	-0.005	0.001
S.E.	(0.011)	(0.011)	(0.024)
R.I. p-value	[0.344]	[0.743]	[0.975]
Difference Between Treatments	-0.020	-0.017	-0.015
S.E.	(0.010)	(0.011)	(0.024)
R.I. p-value	[0.141]	[0.196]	[0.645]
Control Mean	0.945	0.937	0.836
Control SD	0.228	0.244	0.371
N	2252	2251	2249

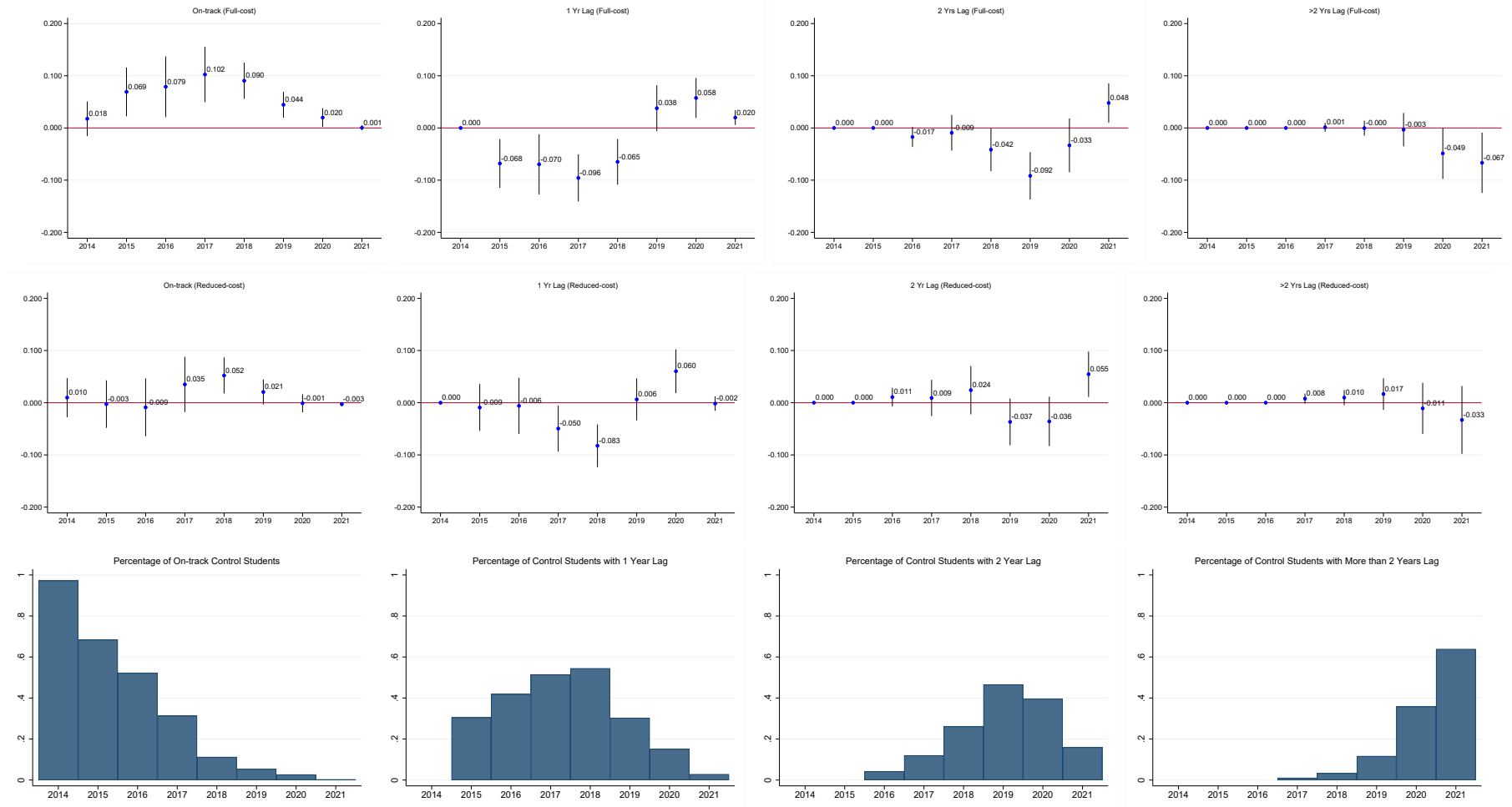
Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference *p*-values, clustered by school (104) and stratified by stratification cell (36), in square brackets [].

level (two in the control group and two in the full-cost version of the program). This was due to the Ugandan government's response to the Covid-19 pandemic, which delayed all grade progression by one year; the tiny number of students observed to be in grade 8 are likely due to measurement error. The fact that grade progression was delayed means that the effect of the NULP on advancing to grade 8 is mechanically zero. Therefore, we use an alternative approach to evaluate whether the program changed students' grade progression.

Figure 5 presents the effects of the NULP on grade progression — on track, one year behind, two years behind or more than two years behind. The rows of the figure presents results for the full-cost, reduced-cost and control group, respectively. Column one shows the probability of being on track with the grade that the student should be in if they progressed one grade level for every school year since 2014. This would mean that we would expect students to be in grade 7 by 2021 (which, as detailed above, essentially never happened). We see that almost all control group students are on-track in 2014 (Grade 1) and thus there are no treatment effects. Already in 2015 only 70% of students in the control group progressed to Grade 2, however we find that significantly more did so in the full-cost group (7 percentage points). This pattern continues up until 2021 were essentially none is on-track. We find no

treatment effects for the reduced-cost version. Column two shows the effect on being one year behind. In 2014 no one is behind and there is no treatment effects. In 2015 30% is one year behind and this is significantly reduced for the full-cost treatment group. After 2019 the probability of being one year behind increases, but at the same time the probability of being two years or more behind decreases (see column three and four). Thus, find that the program is significantly reduces delay in grade progression over time such that students in the full-cost program are fewer years behind compared to what they would have been in absence of the program.

Figure 5
Treatment Effects on Grade Progression



Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Whiskers represent heteroskedasticity-robust 95% confidence intervals, clustered by school.

We also explored additional potential outcomes that could be related to the impact that the program has on the educational outcomes. In particular, we focused on sexual behaviors and labor market participation. Table 10 shows the effects of the program on sexual behavior. By 2021, about 11% of the control students reported having sex at least once (corresponding to 148 students). We find no evidence that either version of the program had any effect on the probability of ever having sex. Among the control students who had had sex, the average age when students had their first sexual experience was about 14 years old. Again, we did not find any evidence of the program having any effect on this variable. Finally, we examined whether the program changed the probability of having sex before the age of 13. Just under 4 percent of the control group had had sex prior to age 13, but we find no statistically significant effects of the NULP on this outcome.

We also explored the possibility that the program affected the probability of the children having a job outside of home. Table 11 presents the results and we find that in 2019, 39% of the control students worked outside of their homes. This percentage grew to 50% and 54% in 2020 and 2021, respectively. Neither version of the program had a statistically significant effect on the probability of working outside of home in any of the three years.

Overall, these results suggest that while the NULP had an impressive and long-lasting effect on Leblango and English learning, as well as a moderate effect in grade progression, these gains did not translate into math learning or other behaviors.

Because our analysis in this paper deviates from our original analysis plan, we present the results based on exactly what we pre-specified in Appendix Section A.3. These results also show the Anderson (2008) FDR-adjusted q-values for the pre-specified outcomes. Appendix Table A9 differs from our main results Table 7 in the exact way that we calculate the test score indices. For our main analysis, we compute the score indices using only those test score components that were available at both baseline and endline, so that our scores can be measured in equivalent years of schooling. In the analysis plan, we pre-specified that we would use all components available at endline. This change makes no difference for our substantive conclusions about the program's effects on test scores. The significance of our estimated treatment effects on test scores is also robust to adjusting for multiple testing. This pattern carries over to Appendix Table A10 and Appendix Table A11: if we run the analyses exactly as pre-specified in our analysis plan, we find no statistically significant effects of the NULP on attending school, attending secondary school, sexual behavior, and working outside the home.

Table 10
Treatment Effects on Sexual Behavior

	(1) Ever Had Sex	(2) Age First Had Sex	(3) Had Sex at 13 or Before
Reduced-cost	-0.010	-0.237	-0.011
S.E.	(0.020)	(0.438)	(0.008)
R.I. p-value	[0.714]	[0.645]	[0.297]
Full-cost	0.006	-0.252	-0.008
S.E.	(0.019)	(0.501)	(0.010)
R.I. p-value	[0.788]	[0.656]	[0.541]
Difference Between Treatments	0.016	-0.015	0.002
S.E.	(0.018)	(0.504)	(0.008)
R.I. p-value	[0.471]	[0.969]	[0.802]
Control Mean	0.109	13.923	0.037
Control SD	0.312	1.929	0.189
N	1690	148	1690

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Column (2) only include students reporting having sex at least once.

Table 11
Treatment Effects on Working Outside of Home

	(1) 2019	(2) 2020	(3) 2021
Reduced-cost	-0.008	0.006	-0.003
S.E.	(0.020)	(0.022)	(0.022)
R.I. p-value	[0.710]	[0.846]	[0.926]
Full-cost	-0.044	-0.030	-0.027
S.E.	(0.023)	(0.024)	(0.020)
R.I. p-value	[0.125]	[0.346]	[0.237]
Difference Between Treatments	-0.036	-0.036	-0.025
S.E.	(0.022)	(0.022)	(0.021)
R.I. p-value	[0.220]	[0.221]	[0.373]
Control Mean	0.393	0.501	0.543
Control SD	0.489	0.500	0.498
N	2228	2231	2227

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assesment, EGMA, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets [].

6 Conclusion

Our findings from this study of the Northern Uganda Literacy Project (NULP) shed light on the sustained effects of education interventions in developing countries. While the literature on such programs primarily focuses on short-term outcomes, this research demonstrates that the NULP had a significant and lasting impact on early grade reading skills, with positive spillovers onto English. Eight years after the program began, a substantial portion of the effects persisted. The study also revealed that the NULP did not have spillover effects onto math or sexual behavior, but did cause improvements in grade progression.

Our findings also highlight dismal overall grade progression in Uganda, even before delays due to school shutdowns during Covid-19. There are limited longitudinal studies that document grade progression (or lack of) in developing countries,[17] and this is an area of important future research.

These findings contribute to the existing literature by highlighting the potential of well-designed early-grade educational interventions for enhancing educational outcomes in developing countries. The sustained effects observed in this study emphasize the importance of sustained efforts and investments in early education programs to support children's learning trajectories—particularly in literacy. Further research is warranted to explore the mechanisms and factors that contribute to the durability of the NULP's effects, as well as to investigate ways to extend its benefits to other academic areas and broader aspects of students' lives.

[17] One exception is Lam, Ardington, and Leibbrandt (2011), which documents very poor grade progression for black students in South Africa.

References

Anderson, Michael L. (2008). "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American statistical Association* 103.484. Publisher: Taylor & Francis, pp. 1481–1495.

Ardington, Cally, Gabrielle Wills, Elizabeth Pretorius, Nompumelelo Mohohlwane, and Alicia Menendez (2021). "Benchmarking oral reading fluency in the early grades in Nguni languages". *International Journal of Educational Development* 84, p. 102433. ISSN: 0738-0593. DOI: 10.1016/j.ijedudev.2021.102433.

Athumai, Halima (2022). "Whatever Happened to the Teens Who Endured the World's Longest COVID School Closure?" en. *NPR*.

Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman, and David S. Yeager (2020). "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions". en. *Psychological Science in the Public Interest* 21.2. Publisher: SAGE Publications Inc, pp. 55–97. ISSN: 1529-1006. DOI: 10.1177/1529100620915848.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel (2016). "Worms at Work: Long-run Impacts of a Child Health Investment". eng. *The Quarterly Journal of Economics* 131.4, pp. 1637–1680. ISSN: 0033-5533. DOI: 10.1093/qje/qjw022.

Bandiera, Oriana, Niklas Buehren, Markus Goldstein, Imran Rasul, and Andrea Smurra (2023). *Safe Spaces for Teenage Girls in a Time of Crisis*. Working Paper.

Barham, Tania, Brachel Champion, Gisella Kagy, and Jena Hamadani (2023). "Improving the Early Childhood Environment: Direct and Distributional Effects on Human Capital for Multiple Generations".

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006). "Adaptive linear step-up procedures that control the false discovery rate". *Biometrika* 93.3. Publisher: Oxford University Press, pp. 491–507.

Bettinger, Eric, Michael Kremer, Maurice Kugler, Carlos Medina, Christian Posso, Juan E Saavedra, et al. (2019). *School vouchers, labor markets and vocational education*. Banco de la Republica Colombia.

Bouguen, Adrien, Yue Huang, Michael Kremer, and Edward Miguel (2019). "Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics". *Annual Review of Economics* 11.1. _eprint: https://doi.org/10.1146/annurev-economics-080218-030333, pp. 523–561. DOI: 10.1146/annurev-economics-080218-030333.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Sebastián Muñoz, Jeffrey Smith, and Rebecca Thornton (2022). "Some Children Left Behind: Variation in the Effects of an Educational Intervention". *Journal of Econometrics* Forthcomming.

Bus, Adriana G. and Marinus H. van IJzendoorn (1999). "Phonological Awareness and Early Reading: A Meta-Analysis of Experimental Training Studies". *Journal of Educational Psychology* 91.3, pp. 403–414. DOI: 10.1037/0022-0663.91.3.403.

Center for Public Education (2015). *Learning to Read, Reading to Learn*. White Paper.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star". en. *The Quarterly Journal of Economics* 126.4, pp. 1593–1660. ISSN: 0033-5533, 1531-4650. DOI: 10.1093/qje/qjr041.

Derksen, Laura, Jason T. Kerwin, Natalia Ordaz Reynoso, and Olivier Sterck (2023). *Healthcare Appointments as Commitment Devices*.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2021). *The Impact of Free Secondary Education: Experimental Evidence from Ghana*. Working Paper 28937. Series: Working Paper Series. National Bureau of Economic Research. DOI: 10.3386/w28937.

Evans, David and Fei Yuan (2019). "Equivalent years of schooling: A metric to communicate learning gains in concrete terms". *World Bank Policy Research Working Paper* 8752.

Evans, David K and Mũthoni Ngatia (2021). "School uniforms, short-run participation, and long-run outcomes: Evidence from Kenya". *The World Bank Economic Review* 35.3. Publisher: Oxford University Press, pp. 705–719.

Evans, David K. and Fei Yuan (2022). "How Big Are Effect Sizes in International Education Studies?" en. *Educational Evaluation and Policy Analysis* 44.3. Publisher: American Educational Research Association, pp. 532–540. ISSN: 0162-3737. DOI: 10.3102/01623737221079646.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeerch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor (2014). "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica". *Science (New York, N.Y.)* 344.6187, pp. 998–1001. ISSN: 0036-8075. DOI: 10.1126/science.1251178.

Hamory, Joan, Edward Miguel, Michael Walker, Michael Kremer, and Sarah Baird (2021). "Twenty-year economic impacts of deworming". *Proceedings of the National Academy of Sciences* 118.14. Publisher: Proceedings of the National Academy of Sciences, e2023185118. DOI: 10.1073/pnas.2023185118.

Hares, Susannah, Jack Rossiter, and Justin Sandefur (2023). *Will Raising Test Scores in Developing Countries Produce More Health, Wealth, and Happiness Later in Life?*

Harvard Graduate School of Education (2016). *Learning to Read to Learn*.

Hattie, John, John Biggs, and Nola Purdie (1996). "Effects of Learning Skills Interventions on Student Learning: A Meta-Analysis". *Review of Educational Research* 66.2, pp. 99–136. ISSN: 0034-6543. DOI: 10.3102/00346543066002099.

Heckman, James, Rodrigo Pinto, and Peter Savelyev (2013). "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes". en. *American Economic Review* 103.6, pp. 2052–2086. ISSN: 0002-8282. DOI: 10.1257/aer.103.6.2052.

Kerwin, Jason T. and Rebecca L. Thornton (2021). "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures". *The Review of Economics and Statistics* 103.2, pp. 251–264. DOI: 10.1162/rest_a_00911.

Lam, David, Cally Ardington, and Murray Leibbrandt (2011). "Schooling as a Lottery: Racial Differences in School Advancement in Urban South Africa". en. *Journal of Development Economics* 95.2, pp. 121–136. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2010.05.005.

Michie, Sammy (2023). "Enabling the Shift from Learning to Read to Reading to Learn". *IMSE - Journal*.

Millán, Teresa Molina, Tania Barham, Karen Macours, John A. Maluccio, and Marco Stampini (2019). "Long-Term Impacts of Conditional Cash Transfers: Review of the Evidence". *World Bank Research Observer* 34.1. Publisher: Oxford University Press, pp. 119–159. ISSN: 1564-6971. DOI: 10.1093/wbro/lky005.

Ozier, Owen (2018). "Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming". en. *American Economic Journal: Applied Economics* 10.3, pp. 235–262. ISSN: 1945-7782. DOI: 10.1257/app.20160183.

Protzko, John (2015). "The Environment in Raising Early Intelligence: A Meta-Analysis of the Fadeout Effect". *Intelligence* 53, pp. 202–210. ISSN: 0160-2896. DOI: 10.1016/j.intell.2015.10.006.

Stern, Jonathan M B, Matthew C H Jukes, Jacobus Cilliers, Brahm Fleisch, Stephen Taylor, and Nompumelelo Mohohlwane (2023). *Persistence and Emergence of Literacy Skills: Long-Term Impacts of an Effective Early Grade Reading Intervention in South Africa*. Tech. rep. Working Paper 672. Center for Global Development.

Takacs, Zsofia K. and Reka Kassai (2019). "The Efficacy of Different Interventions to Foster Children's Executive Function Skills: A Series of Meta-Analyses". eng. *Psychological Bulletin* 145.7, pp. 653–697. ISSN: 1939-1455. DOI: 10.1037/bul0000195.

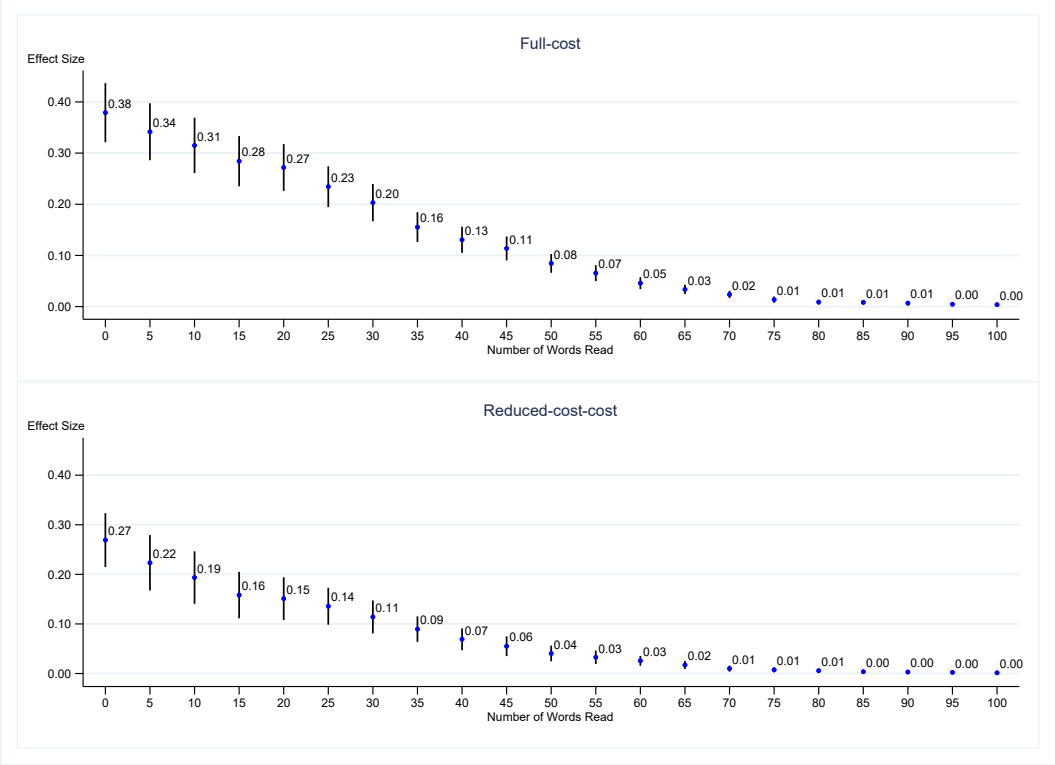
Uwezo (2016). *Are Our Children Learning (2016)? Uwezo Uganda Sixth Learning Assessment Report*. Tech. rep. Kampala: Twaweza East Africa.

World Bank (2020). "School enrollment, primary (% net)". *World Development Indicators*.

A Online Appendix

A.1 Appendix Figures

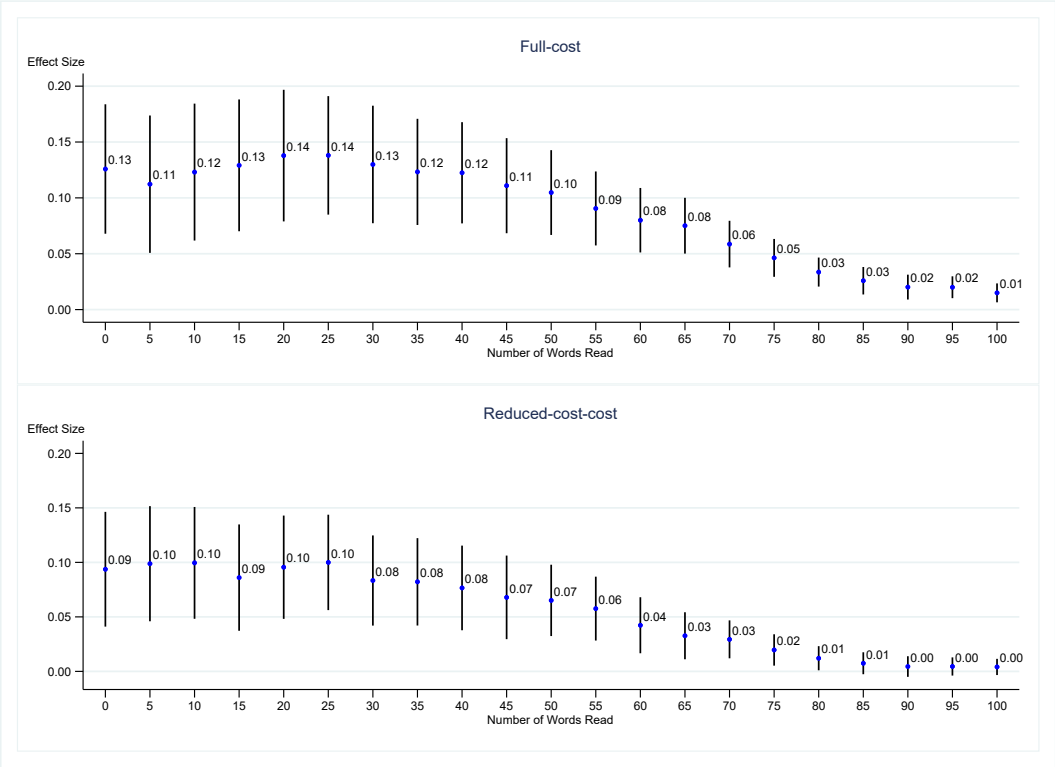
Appendix Figure A1
Immediate Impacts on Leblango Words Read per Minute



Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Whiskers represent heteroskedasticity-robust 95% confidence intervals, clustered by school.

Appendix Figure A2

Treatment Effects on Leblango Words Read per Minute (Five Year Follow-up)



Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Whiskers represent heteroskedasticity-robust 95% confidence intervals, clustered by school.

A.2 Appendix Tables

Appendix Table A1

Grade Attended by Control Students each Year, Males

Grade Attended	Year							
	2014	2015	2016	2017	2018	2019	2020	2021
Not enrolled	0.02	0.01	0.02	0.04	0.05	0.06	0.05	0.17
1	0.98	0.33	0.05	0.02	0.00	0.00	0.00	0.00
2	0.00	0.66	0.42	0.13	0.03	0.02	0.01	0.00
3	0.00	0.00	0.51	0.53	0.28	0.13	0.06	0.04
4	0.00	0.00	0.00	0.28	0.55	0.46	0.34	0.26
5	0.00	0.00	0.00	0.00	0.09	0.28	0.40	0.36
6	0.00	0.00	0.00	0.00	0.00	0.05	0.13	0.15
7	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

Appendix Table A2

Grade Attended by Control Students each Year, Females

Grade Attended	Year							
	2014	2015	2016	2017	2018	2019	2020	2021
Not enrolled	0.03	0.01	0.01	0.04	0.04	0.05	0.08	0.16
1	0.96	0.29	0.04	0.01	0.00	0.00	0.00	0.00
2	0.01	0.69	0.42	0.11	0.03	0.00	0.00	0.00
3	0.00	0.01	0.52	0.51	0.25	0.08	0.04	0.03
4	0.00	0.00	0.01	0.33	0.55	0.49	0.27	0.25
5	0.00	0.00	0.00	0.00	0.12	0.33	0.41	0.36
6	0.00	0.00	0.00	0.00	0.00	0.05	0.17	0.17
7	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01

Notes: Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

Appendix Table A3

Grade Attended by Control Students each Year, Students Scoring Higher than Zero in Leblango EGRA at the Beginning of 2014

Grade Attended	Year							
	2014	2015	2016	2017	2018	2019	2020	2021
Not enrolled	0.01	0.02	0.02	0.05	0.04	0.04	0.07	0.15
1	0.98	0.20	0.02	0.01	0.00	0.00	0.00	0.00
2	0.01	0.77	0.33	0.08	0.02	0.01	0.00	0.00
3	0.00	0.01	0.62	0.43	0.17	0.04	0.03	0.02
4	0.00	0.00	0.01	0.43	0.56	0.40	0.17	0.13
5	0.00	0.00	0.00	0.01	0.20	0.40	0.44	0.40
6	0.00	0.00	0.00	0.00	0.01	0.11	0.23	0.22
7	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.06
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01

Notes: Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

Appendix Table A4

Grade Attended by Control Students each Year, Students Scoring Zero in Leblango EGRA at the Beginning of 2014

Grade Attended	Year							
	2014	2015	2016	2017	2018	2019	2020	2021
Not enrolled	0.03	0.00	0.01	0.04	0.04	0.06	0.07	0.15
1	0.97	0.32	0.04	0.01	0.00	0.00	0.00	0.00
2	0.00	0.68	0.44	0.11	0.02	0.01	0.00	0.00
3	0.00	0.00	0.50	0.55	0.28	0.10	0.05	0.04
4	0.00	0.00	0.00	0.29	0.56	0.50	0.32	0.27
5	0.00	0.00	0.00	0.00	0.09	0.29	0.41	0.36
6	0.00	0.00	0.00	0.00	0.00	0.04	0.13	0.16
7	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year. Grades 1 to 7 is primary school and Grades 8 and 9 are secondary school.

Appendix Table A5
Grade Attended by Control Students as Reported by Schools

Grade Attended	Year			
	2014	2015	2016	2017
1	1.00	0.00	0.00	0.00
2	0.00	1.00	0.01	0.00
3	0.00	0.00	0.99	0.01
4	0.00	0.00	0.00	0.99
5	0.00	0.00	0.00	0.00

Notes: This table shows which grade control students were attending according to their schools between 2014 and 2017. Bordered cells reference which grade the students are expected to be in if they are on track. Grey cells show the modal grade each year.

Appendix Table A6
Grade Repetition by Control Students each Year

	Have you repeated the grade you attended that year?	
Year	No	Yes
2014	0.98	0.02
2015	0.70	0.30
2016	0.80	0.20
2017	0.70	0.30
2018	0.62	0.38
2019	0.54	0.46
2020	0.53	0.47
2021	0.75	0.25
Ever repeated	0.07	0.93

Notes: This table shows the percentage of control students that reported in 2021 they had repeated the grade they were attending each year.

Appendix Table A7
Grades Repeated by Control Students

Grade	Students
1	0.35
2	0.26
3	0.45
4	0.72
5	0.26
6	0.03
7	0.00
8	0.00
9	0.00
Never repeated	0.07

Notes: This table shows the percentage of control students that repeated each grade as reported by students in 2021.

Appendix Table A8
Number of Grades Repeated by Control Students

Number of grades repeated	% of students
0	0.07
1	0.20
2	0.36
3	0.25
4	0.11
5	0.01

Notes: This table shows how many grades control students repeated as reported by students in 2021.

A.3 Analysis Plan Results

This appendix shows the estimations of the treatment effects of the NULP that mirror the preanalysis plan. These results don't change the conclusions presented in Section 5.

Appendix Table A9
Treatment Effects on Confirmatory Academic Outcomes

	English	Leblango	Math
Reduced-cost	0.244***	0.377***	-0.027
S.E.	(0.099)	(0.148)	(0.048)
R.I. p-value	[0.001]	[0.000]	[0.662]
q-value	{0.002}	{0.001}	{0.284}
Full-cost	0.537***	0.712***	0.002
S.E.	(0.138)	(0.200)	(0.044)
R.I. p-value	[0.001]	[0.000]	[0.970]
q-value	{0.002}	{0.001}	{0.478}
Difference Between Treatments	0.293	0.335	0.028
S.E.	(0.150)	(0.213)	(0.043)
R.I. p-value	[0.118]	[0.237]	[0.589]
q-value	{0.548}	{0.548}	{0.552}
Control Mean	-0.000	-0.000	-0.000
Control SD	1.000	1.000	1.000
N	3098	3098	3098

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Multiple testing adjusted q-values in curly braces {}. Coefficients represent standard deviations of the control group.

Appendix Table A10

Treatment Effects on Confirmatory Downstream Outcomes

	Attended School (2021)	Attended Secondary School (2021)
Reduced-cost	0.016	-0.003
S.E.	(0.025)	(0.002)
R.I. p-value	[0.575]	[0.424]
q-value	{0.935}	{0.935}
Full-cost	0.001	0.001
S.E.	(0.024)	(0.002)
R.I. p-value	[0.975]	[0.925]
q-value	{0.935}	{0.935}
Difference Between Treatments	-0.015	0.003
S.E.	(0.024)	(0.002)
R.I. p-value	[0.645]	[0.176]
q-value	{0.583}	{0.583}
Control Mean	0.836	0.003
Control SD	0.371	0.052
N	2249	2249

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Multiple testing adjusted q-values in curly braces {}.

Appendix Table A11
Treatment Effects on Exploratory Outcomes

	Ever had Sex	First had sex at age 13 or below	Worked outside of the home
CCT	-0.010	-0.011	-0.003
S.E.	(0.020)	(0.008)	(0.022)
R.I. p-value	[0.714]	[0.297]	[0.926]
q-value	{1.000}	{1.000}	{1.000}
MT	0.006	-0.008	-0.027
S.E.	(0.019)	(0.010)	(0.020)
R.I. p-value	[0.788]	[0.541]	[0.237]
q-value	{1.000}	{1.000}	{1.000}
Difference Between Treatments	0.016	0.002	-0.025
S.E.	(0.018)	(0.008)	(0.021)
R.I. p-value	[0.471]	[0.802]	[0.373]
q-value	{1.000}	{1.000}	{1.000}
Control Mean	0.109	0.037	0.543
Control SD	0.312	0.189	0.498
N	1690	1690	2227

Notes: All regressions control for stratification cell indicators and baseline values of the Leblango EGRA Score, Oral English assessment, math test, gender and age; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). Randomization inference p-values, clustered by school (104) and stratified by stratification cell (36), in square brackets []. Multiple testing adjusted q-values in curly braces {}.