Online Appendix to Making the Grade: The Trade-off between Efficiency and Effectiveness in Improving Student Learning

Jason T. Kerwin and Rebecca L. Thornton

July 13, 2017

Click here for the latest version of this appendix

A Intervention Inputs

The full-cost and reduced-cost 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 Appendix Table A1.

A.1 Materials

The NULP provides the following materials to each full-cost and reduced-cost 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 full-cost program provides additional materials to each school:

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

A.2 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 (full-cost program only)
- Three trainings in literacy methods for P1-P3 teachers during the school holidays each year
    - Full-cost program: residential trainings held in the district capital, conducted by experienced MT staff
    - Reduced-cost program: non-residential trainings held at the CCs, conducted by CCTs.
- Special field monitoring and support supervision visits to schools
    - Full-cost program: 3 times per term by project staff, 2 times per term by CCTs
    - Reduced-cost program: 2 times per term by CCTs

A.3 Other Support

The NULP also provides other support to schools and communities:

- Schools in both the full-cost program and the reduced-cost 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 full-cost program schools and the CCTs for the reduced-cost 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 is a simple way for parents to assess their children in basic reading skills. [1]
    - The tool has 4 parts: 1) letter name knowledge; 2) familiar word reading; 3) reading fluency test; and 4) reading comprehension test.
- Mango Tree sponsors a one-hour monthly radio program (supported by SMS messages and surveys to engage listeners in feedback) that broadcasts literacy and

[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.

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.

- (Full-cost 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.

A.4 Arancibia, Popova, and Evans (2016) Indicators

As an objective, third-party summary the differences between study arms, we use the coding scheme developed by Arancibia, Popova, and Evans (2016). This is a tool for recording the attributes of in-service teacher training programs. Appendix Table A2 presents the way in which they coded the attributes of the two NULP variants, based on an interview with a member of the Mango Tree team. Excluding the three indicators that measure sample size, there are just three indicators (out of a total of 48) that differ between the two versions of the program: whether the program was a cascade training model (0=no for full-cost, 1=yes for reduced-cost), the profile of the trainers (1=primary or secondary teachers for full-cost, 4=local government official for reduced-cost) and the number of follow-up visits (9 for full-cost, 6 for reduced-cost). This follow-up visit figure understates the difference across study arms somewhat: full-cost schools receive 3 visits per term from Mango Tree staff on top of the 2 visits per term from CCTs, for a total of 15 visits rather than 9. This would still be coded as a difference, however, and so does not change our conclusions about how different the two versions of the program are.

B Baseline Balance

Appendix Table A3 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

$$y_{is} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{L}_s'\gamma + \epsilon_{is} \qquad \text{(A1)}$$

Here we control for stratification cell indicators $\boldsymbol{L}_s'$ because the fraction of schools in each study arm varies by cell. We cluster our standard errors at the school level. Statistically significant differences are indicated by stars next to the full-cost and reduced-cost 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; there are some differences in individual exam components that wash out when we look at each test as a whole. 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 means 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. 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, but is balanced across study arms in terms of the overall test score indices. While the overall test scores are balanced, the individual test components do exhibit slight imbalance for certain comparisons. However, the patterns of statistically-significant coefficients are consistent with what would be expected by random chance. To address potential concerns about imbalance, we control for baseline test scores in our preferred regression specification (equation 1).

C Classroom Observation Data Details

C.1 Data Collection Instrument

The classroom observation instrument consists of two parts. Part 1 records metadata such as the school and teacher being observed, the enumerator name and the date of the survey. Part 2 contains three identical pages, one for each ten-minute observation window in the thirty-minute class being observed. Appendix Figure A1 shows an example of one of these pages. It allows the enumerator to record teacher and student actions that take place during each observation window.

C.2 Factor Loadings for Classroom Observation Data

Appendix Tables A4 through A6 show the rotated factor loadings from the exploratory factor analyses we conducted on the classroom management variables, the reading pedagogy variables, and the writing pedagogy variables respectively. The names for each factor index are descriptive terms we created based on which variables the indices load most heavily on. We also present the share of the variance of the data explained by each factor.

D Robustness Checks

D.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 2 and 3, but modifying equation (1) to remove these controls for baseline values. Our modified regression equation is:

$$y_{is} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{L}_s'\gamma + \epsilon_{is} \qquad \text{(A2)}$$

Here *i* indexes students and s indexes schools. y_{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. The results are presented in Appendix Tables 7 and 8, which mirror Tables 1 and 2 in the main text. The point estimates and standard errors are nearly unaffected by the exclusion of the controls. For the EGRA (Appendix Table 7), the regression without baseline test score controls yields slightly larger effect sizes for the full-cost program and slightly smaller effect sizes for the reduced-cost program.

For the writing test (Appendix Table 8), omitting the baseline test score controls leads to marginally smaller estimated the gains for students in the full-cost variant of the program, and marginally larger estimated losses on advanced writing skills (and overall writing ability) for students in the reduced-cost version.[2]

[2] Column 10 is identical between Table T_Writing_Results and Appendix Table 3 because Presentation was not one of the scored categories at baseline. Columns 6 (Voice) and 9 (Conventions) are also identical because no pupils

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

D.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 identifying assumption that treatment status is 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 A9 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 full-cost version of the program are similar but somewhat larger; the combined PCA index shows a 50% larger increase using the reduced sample. For the reduced-cost 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 at the 5% level, as compared with three for the full sample. Overall, the results are not particularly sensitive to the inclusion of this stratification cell.

3.2 Effect of NULP on Oral English Speaking Ability

---

received any points for those categories at baseline, so the controls were dropped due to collinearity with the constant term.

The eventual goal of both the standard government curriculum and the NULP model is for students to successfully transition to English after P3. One potential question about mother tongue education is whether it crowds out the ability to speak, read, and write, in English. We administered a simple oral English examination – designed by Mango Tree – that measures basic speaken English vocabulary using pictures. The oral English examination has three sections. The first section 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 and presents students with a picture of a scene and asks them what objects and which people they can see in the picture.[3]

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 Mango Tree model contrasts sharply with that practice, and explicitly instructs teachers to avoid using any written English text during P1.

Appendix Table 10 presents the effects of the two program variants on students' scores on the oral English examination, estimated using equation (1). Neither the full-cost nor the reduced-cost version of the program has 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 full-cost version raises this index by 0.14 SDs, and the reduced-cost 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 reduced-cost program but mostly positive for the full-cost version. Moreover, Columns 8 and 9 show that the full-cost program had statistically-

[3] 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).

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 Lango 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 full-cost 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 Lango, 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, in the *status quo* it is common in classrooms in the Lango 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 reduced-cost version and by 0.29 SDs under the full-cost 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.

Appendix Figure A1
Classroom Observation Instrument

Specific Lesson Actions *(Repeated for Second and Third Ten-Minute Window)*

<table>
<tr><th>Time</th><th>Teacher actions</th><th colspan="4">Pupil actions</th></tr>
<tr><td rowspan="8">FIRST
10 minutes:

_____
(start time)

_____
(end time)</td><td rowspan="8">Positive actions:
□ Refers to TG or lesson plan while teaching
□ Moves freely around the classroom
□ Calls on individual pupils by name
□ Encourages pupil participation and keeps their attention
□ Brings pupils back on task when needed
□ Observes and records pupils' performance

Negative actions:
□ Lesson does not appear planned
□ Remains at the front of the class
□ Does not call on individual pupils by name
□ Very little pupil participation and attention
□ Ignores or does not address pupils who are off task
□ Does not record pupil performance

Other:
% time speaking English _____%
% time speaking LL _____%
Minutes out of class _____ min.
Minutes in class but not teaching _____ min.
Minutes teaching _____ min.</td><td colspan="4">Reading</td></tr>
<tr><td>□ Sounds
□ Letters
□ Words
□ Sentences</td><td>□ Whole class
□ Smaller group
□ Individual at seat
□ Individual at board</td><td>□ On board
□ In primer
□ In reader
□ Other:
_____</td><td>□ English
□ LL</td></tr>
<tr><td colspan="4">Minutes on pupil reading tasks _____ min.
% of pupils participating in reading task _____%</td></tr>
<tr><td colspan="4">Writing</td></tr>
<tr><td>□ Pictures
□ Letters
□ Words
□ Sentences
□ Name</td><td>□ Air writing
□ Handwriting practice
□ Copying teacher text from the board
□ Writing own text</td><td>□ On slate
□ On paper
□ On board</td><td>□ English
□ LL</td></tr>
<tr><td colspan="4">Minutes on pupil writing tasks _____ min.
% of pupils participating in writing task _____%</td></tr>
<tr><td colspan="4">Speaking/Listening</td></tr>
<tr><td></td><td>□ To a partner
□ To a small group
□ To the whole class
□ To the teacher</td><td></td><td>□ English
□ LL</td></tr>
<tr><td colspan="4">Minutes on pupil speaking/listening tasks _____ min.
% of pupils participating in speaking/listening task _____%</td></tr>
</table>

Appendix Table A1

NULP Components by Study Arm

	Study Arm		
	Full-cost program	Reduced-cost program	Control
Number of Schools	12	14	12
Pedagogy			
Local Language-First Instruction	Yes	Yes	
NULP Instructional Model	Yes	Yes	
Books			
Leblango Primers	3 per student (1 for each term)	3 per student (1 for each term)	
Leblango Readers	3 per student (1 for each term)	3 per student (1 for each term)	
Leblango Teacher's Guides	1 per classroom	1 per classroom	
English Primers	3 per student (1 for each term)	3 per student (1 for each term)	
English Teacher's Guides	1 per classroom	1 per classroom	
Materials			
Slates	1 per student		
Wall Clocks	1 per classroom		
Training and Support for Teachers			
Leblango Orthography Training (5 Days, before school year)	Before school year, non-residential, taught by MT staff	Before school year, residential, taught by CCTs	
Literacy Methods Training (3-5 days, before each term)	1X/term, residential, taught by MT staff	1X/term, non-residential, taught by CCTs	
Saturday in-service training wkshps (1 Day, during each term)	2X/term, non-residential, taught by MT staff	2X/term, non-residential, taught by CCTs	
Classroom support supervision	3X/term from MT staff, 2X/term from CCTs	2X/term from CCTs	
Other			
Take a Book Home Activity	Early during first term		
Literacy & Local Language Radio Program	1X/month, available to whole community		

Appendix Table A2

Comparison of Arancibia, Popova, and Evans (2016) Indicators for Full-Cost and Reduced-Cost NULP

	Full-Cost	Reduced-Cost
Which organization designed the program?	2	2
Which organization is implementing the program?	2	2
Was program design based on a diagnostic or evaluation of some kind? If so, which one?	1	1
Program objectives	To create a culture of literacy and engage with people responsible for growing this culture. For students to learn the names of the letters of the	
Targeting by geography	1	1
Targeting by subject	0	0
Targeting by grade	1	1
Targeting by years of experience	0	0
Targeting by skill gaps	0	0
Targeting by contract teachers	0	0
Do teachers have to pay some cost for the training (including their own transport cost)? If so, how much over one school year?	0	0
Does participation have any of these implications?	0	0
Is there a positive consequence if teachers are well evaluated?	0	0
Is there a negative consequence if teachers are poorly evaluated?	0	0
Did the program provide textbooks?	0	0
Did the program provide storybooks?	1	1
Did the program provide computers?	0	0
Did the program provide teacher manuals?	1	1
Did the program provide lesson plans/videos?	1	1
Did the program provide scripted lessons?	1	1
Did the program provide craft materials?	0	0
Did the program provide other reading materials - flashcards, word banks, reading pamphlets or similar?	1	1
Did the program provide software?	0	0
How many teachers received training under this program each year?	24	28
How many schools is the program being implemented in (at the time of the evaluation)?	12	14
How many years has the program been running (at the time of the evaluation)?	2	2
In the last year what percentage of the teachers who began the training dropped out before the end?	0	0
What is the primary focus of the training program?	2	2
What is the secondary focus of the training program?	1	1
What is the subject focus of the training program (if any)?	1	1
Did the training involve lectures?	1	1
Did the training involve discussion?	1	1
Did the training involve lesson enactment?	1	1
Did the training involve materials development?	0	0
Did the training involve training on how to conduct diagnostics?	1	1
Did the training involve lesson planning?	1	1
Did the training involve the use of scripted lessons?	1	1
Is it a cascade training model (i.e. one where program trainers trainers who then train teachers)?	0	1
What is the most common profile of the direct trainers?	1	4
Is there a part of the training where teachers meet with trainers for several days in a row?	1	1
During this period, what is the total hours of teacher training they receive?	120	120
During this period, how many hours of lectures do they receive?	60	60
During this period, how many hours do they spend practicing with students?	0	0
During this period, how many hours do they spend practicing with other teachers?	60	60
Over how many weeks?	40	40
Where does this part of the training take place?	2	2
How many teachers are in each training session?	24	26
How many in-school follow-up support visits do teachers receive after the initial training (if any)?	9	6
What is the nature of these follow-up visits?	1	1
How many weeks of distance learning does the program include (if any)?	0	0
Over how many months?	9	9
Tested subject	Average	Average
Africa dummy	1	1
Interviewed	1	1

Appendix Table A3

Baseline Covariate Means by Study Arm

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Baseline Sample			Longitudinal Sample			Lost to Followup		
	Control	Full-Cost	Reduced-Cost	Control	Full-Cost	Reduced-Cost	Control	Full-Cost	Reduced-Cost
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	-0.075*	0.001	0.046	-0.100	-0.003	-0.160	-0.038
1(any correct)	0.396	0.386	0.368	0.394	0.406	0.378	0.402	0.301*	0.341
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
Writing Test									
PCA writing score index	0.000	-0.011	-0.027	0.067	0.001	-0.144	-0.259	-0.130*	-0.226
1(any correct)	0.212	0.330*	0.186	0.237	0.355**	0.195	0.114	0.226	0.160
African name (surname) writing	0.180	0.323**	0.181	0.201	0.348***	0.193	0.098	0.217**	0.149
English name (given name) writing	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.

Appendix Table A4

Factor Loadings for Classroom Management Indices

	(1) Keeps Students Focused	(2) Solid Lesson Plan	(3) Active Throughout Classroom
Teacher Actions:			
Refers to Teacher's Guide	0.01	0.34	0.05
Moves Freely Around Classroom	0.00	-0.03	0.32
Calls on Individuals	0.02	0.09	0.13
Brings Students Back on Task	0.48	-0.01	0.13
Observes/ Records Performance	0.02	0.07	0.27
Lesson Not Planned	0.01	-0.31	0.05
Very Little Participation	-0.06	-0.13	-0.01
Ignores Off-Task Students	-0.42	0.06	0.19
Share of Time Speaking Leblango	-0.02	-0.03	-0.06
Share of Variance Explained	0.81	0.31	0.25

Notes: This table presents the rotated factor loadings for the three indices of classroom management techniques used in the paper. We retain all factors that explain at least 10% of the variance of the data, and apply a varimax rotation to the resulting set of selected factors. We then give each factor a descriptive name based on which of the underlying behaviors it loads on.

Appendix Table A5

Factor Loadings for Reading Pedagogy Indices

	(1)	(2)	(3)	(4)	(5)
	Sounds and Letters Only	Whole Language On Board	Basic Elements in Breakout Sessions	Leblango Sentences in Reader	Paragraphs in Primer
Students are Reading:					
Sounds	0.27	0.01	-0.02	0.07	0.10
Letters	0.41	0.04	0.01	0.09	0.01
Words	0.01	0.05	0.17	-0.10	-0.02
Sentences	-0.29	0.08	-0.02	0.25	0.14
Whole Paragraphs	0.00	0.03	0.16	-0.06	0.14
In Smaller Groups	-0.05	0.06	0.26	-0.01	-0.02
Individually at Seats	0.03	0.02	0.27	0.07	0.03
Individually on Board	-0.03	0.08	-0.06	0.07	-0.17
Whole Group on Board	0.01	0.52	0.00	-0.02	0.06
In Primer	0.00	-0.20	-0.05	-0.05	0.27
In Reader	0.03	-0.13	0.14	0.24	-0.13
From Other Text	-0.04	-0.06	0.17	-0.10	-0.18
Percent of Students Participating	0.02	-0.03	0.08	-0.02	0.16
Share in Leblango	0.03	0.02	0.04	0.29	-0.01
Share of Variance Explained	0.49	0.35	0.27	0.19	0.15

Notes: This table presents the rotated factor loadings for the five indices of reading pedagogy used in the paper. We retain all factors that explain at least 10% of the variance of the data, and apply a varimax rotation to the resulting set of selected factors. We then give each factor a descriptive name based on which of the underlying behaviors it loads on.

Appendix Table A6

Factor Loadings for Writing Pedagogy Indices

	(1) Pictures, Words, and Stories	(2) Copying Teacher's Text	(3) Leblango Practice on Slates	(4) Pictures and Letters on Paper, High-Energy	(5) Leblango Sentences and Handwriting
Students are Writing:					
Pictures	0.15	-0.04	0.11	0.12	-0.14
Letters	-0.50	0.04	0.15	0.11	-0.08
Words	0.10	0.11	0.04	-0.07	-0.04
Sentences	0.04	0.05	-0.02	0.03	0.34
Their Names	0.06	0.00	0.24	0.00	0.07
Air Writing	-0.22	-0.13	0.00	-0.05	0.04
Handwriting Practice	-0.01	0.02	0.15	0.03	0.26
Copying Teacher's Text from Board	0.05	0.44	0.09	0.03	-0.04
Writing Own Text	0.12	-0.34	0.08	0.07	-0.07
On Slate	0.01	0.00	0.31	-0.11	-0.03
On Paper	0.06	0.06	-0.11	0.39	0.04
On Board	0.00	-0.02	-0.11	-0.22	-0.01
Percent of Students Participating	-0.01	0.01	0.08	0.14	-0.12
Share in Leblango	-0.05	-0.06	0.18	0.01	0.11
Share of Variance Explained	0.46	0.31	0.21	0.16	0.12

Notes: This table presents the rotated factor loadings for the five indices of writing pedagogy used in the paper. We retain all factors that explain at least 10% of the variance of the data, and apply a varimax rotation to the resulting set of selected factors. We then give each factor a descriptive name based on which of the underlying behaviors it loads on.

Appendix Table A7

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 Leblango EGRA Score Index†	Letter Name Knowledge	Initial Sound Recogniton	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Full-cost Program	0.654***	1.043***	0.649***	0.382***	0.233**	0.484***	0.449***
	(0.127)	(0.163)	(0.129)	(0.091)	(0.097)	(0.121)	(0.110)
Reduced-cost Program	0.110	0.418**	0.064	-0.012	0.021	0.058	0.034
	(0.102)	(0.181)	(0.096)	(0.074)	(0.069)	(0.081)	(0.084)
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	0.096	0.056	0.037	0.063	0.051
Difference between full-cost and reduced-cost treatment effects	0.544***	0.624***	0.585***	0.393***	0.213**	0.426***	0.415***
	(0.124)	(0.159)	(0.127)	(0.092)	(0.093)	(0.115)	(0.12)
Raw (unstandardized) scores							
Control Group Mean§	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control Group SD§	1.000	9.364	1.92	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table A8

Program Impacts on Writing Test Scores, 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 Writing Score Index†	African Name (Surname) Writing	English Name (Given Name) Writing	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conven-tions	Presen-tation
Full-cost 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)
Reduced-cost 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
Adjusted R-Squared	0.265	0.193	0.217	0.161	0.304	0.177	0.165	0.3	0.164	0.171
Difference between full-cost and reduced-cost treatment effects	0.631***	0.577***	0.837***	0.435***	0.758***	0.465***	0.436***	0.711***	0.474***	0.469***
	(0.149)	(0.136)	(0.156)	(0.151)	(0.173)	(0.118)	(0.15)	(0.175)	(0.151)	(0.115)
Raw (unstandardized) scores										
Control Group Mean§	0.482	0.593	2.000	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control Group SD§	1.000	2.000	0.533	0.372	0.594	0.393	0.416	0.59	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table A9

Program Impacts on Writing Test Scores, Excluding Stratification Cell for School That Completed Test 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
Full-cost Program	0.613***	0.933***	1.364***	0.372***	0.701***	0.350***	0.351***	0.638***	0.435***	0.328***
	(0.108)	(0.117)	(0.150)	(0.109)	(0.129)	(0.091)	(0.114)	(0.130)	(0.110)	(0.088)
Reduced-cost Program	-0.004	0.473***	0.527***	-0.093	-0.079	-0.130**	-0.107	-0.093	-0.050	-0.155**
	(0.076)	(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
Adjusted R-Squared	0.315	0.234	0.241	0.153	0.319	0.165	0.151	0.302	0.146	0.158
Difference between full-cost and reduced-cost treatment effects	0.618***	0.46***	0.837***	0.464***	0.78***	0.48***	0.458***	0.731***	0.485***	0.484***
	(0.117)	(0.144)	(0.162)	(0.130)	(0.146)	(0.091)	(0.127)	(0.147)	(0.130)	(0.090)
Raw (unstandardized) scores										
Control Group Mean§	0.222	0.527	0.274	0.061	0.131	0.084	0.075	0.108	0.037	0.098
Control Group SD§	0.585	0.671	0.486	0.239	0.338	0.278	0.264	0.31	0.19	0.298

Notes: Sample includes 1,478 students from 38 schools 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 Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table A10

Program Impacts on Oral English Test Scores & English Word Recognition

(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Oral English Score Index†	(2) Test 1 (Vocab.)	(3) Test 1 (Count)	(4) Test 2a (Vocab.)	(5) Test 2a (Phrase Structure)	(6) Test 2b (Vocab.)	(7) Test 2b (Phrase Structure)	(8) Test 3 (Vocab., Expressive: Objects)	(9) Test 3 (Vocab., Expressive: People)	(10) Reading English Words‡
Full-cost program	0.145	0.157	-0.118	-0.034	0.045	0.025	-0.114	0.306***	0.295**	-0.290**
	(0.099)	(0.099)	(0.097)	(0.095)	(0.114)	(0.100)	(0.113)	(0.105)	(0.117)	(0.135)
Reduced-cost program	-0.087	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
Adjusted R-Squared	0.346	0.164	0.163	0.205	0.186	0.279	0.092	0.238	0.188	0.274
Difference between full-cost and reduced-cost treatment effects	0.233**	0.156	-0.002	-0.014	0.158*	0.179*	0.098	0.330***	0.394***	-0.080
	(0.092)	(0.099)	(0.072)	(0.092)	(0.089)	(0.092)	(0.092)	(0.104)	(0.093)	(0.108)
Raw (unadjusted) values§										
Control Group Mean	0.101	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	0.62	0.911	1.209	1.928	2.217	2.133	1.839	4.184

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Recognition of Printed English Words (column 10), which was not administered at baseline; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA Oral English Score Index is constructed by weighting each of the 8 test modules (columns 2 through 9) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006). The index is normalized by subtracting the baseline control-group mean and dividing by the endline control-group standard deviation, so that the control group mean for the index shows the control group's progress over the course of the year. Estimated effects are comparable for an alternative index that uses the unweighted mean across (normalized) test modules instead.

‡ Reading English Words is not part of the Oral English examination (and is not included in the computation of the overall PCA index).

§ Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.