Online Appendix to

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

Jason T. Kerwin and Rebecca L. Thornton

January 29, 2020

Click here for the latest version of this appendix

Appendix A: School Eligibility Criteria

To be eligible for the RCT, schools had to meet the following criteria:

a) two first-grade classrooms and teachers

b) desks and lockable cabinets in each classroom

c) a student-teacher ratio less than 135 during 2012 in grades one to three

d) located less than 20 km from the Coordinating Centre headquarters

e) accessible by road; f) a head teacher regarded as "engaged" by the CCT

g) no previous Mango Tree support.

Prior to treatment assignment, head teachers were asked to assign their two best teachers to first-grade and sign a contract with Mango Tree outlining guidelines for study participation. In previous years, while the program was being piloted, schools that did not adhere to the contracts lost Mango Tree support.

Appendix B: Statistical Power

We conducted *ex ante* power calculations using Mango Tree's records from their work piloting the NULP in a separate set of schools from the ones used in the RCT. Mango Tree had conducted Leblango EGRA exams at the beginning and end of 2010 to 2012 in pilot schools. In 2012, Mango Tree internal evaluators also choose a sample of "comparable" non-program schools that they viewed as similar to their program schools. Using these data, we estimated an effect of 1.6 SDs on letter name recognition. Our power calculations, specified in our pre-analysis plan, indicated that our minimum detectable effects (MDE) would be 0.33 SDs (80% power, 0.20 ICC, comparing 12 schools to another 12 schools).[1]

We can also conduct post-hoc power analyses following Ioannidis et al. (2017), by using our estimated standard error to determine the MDE. Post-hoc power calculations that use the estimated treatment effect are subject to type-M error and tend to show that any study with a statistically-significant treatment effect is well-powered (Gelman and Carlin 2014). McKenzie and Ozier (2019) show that using the estimated standard error to construct a MDE does not have the same issue: data generated under a DGP with a given true MDE will have an estimated MDE that is close to the true value, even if only datasets with statistically-significant treatment effects are used.

[1] Our initial calculations assumed 145 students per school, but our final sample averages just 38 students per school. We estimated a partial R-squared for past test scores of 0.7 based on the year-on-year predictive power of test scores for older students; our actual R-squared is just 0.04 because most students initially cannot read at all. The observed ICC in our data is 0.16. If we use these values instead, our MDE at 80% power is 0.50 SDs with 80% power.

The MDE for 80% power is 2.8 times the standard error, or 0.38 SDs for the effect of the full-cost program on overall reading, 0.47 SDs for letter name knowledge, and 0.40 SDs for overall writing. Standard power calculators do not correct for a small number of clusters, and our corrections for our small numbers of clusters produce only *p*-values and not standard errors. Instead, following Ioannidis et al. (2017), we can take the half-width of the 99.5% confidence interval as an estimate of the MDE at 80% power—the same cutoff that is selected using the 2.8 standard error rule. Using the boottest command (Roodman et al. 2019), we find MDEs at 80% power of 0.64 SDs for overall reading, 0.79 SDs for letter name knowledge, and 0.62 SDs for overall writing.

Appendix C: Factor Analyses

In addition to analyzing the raw classroom observation variables, we also conduct factor analyses following Glewwe, Ross, and Wydick (2018). This approach lets us summarize the patterns of correlations between different variables in the classroom observations. Using the raw variables that measure teacher and student behaviors, material use, and time allocation, we conduct factor analyses separately for classroom management, reading activities, and writing activities. Because the treatment may alter the patterns of behavior in the classroom, we use data from all three study arms to estimate the factors. We retain all factors that explain at least 10% of the variance in the data and give descriptive names to each factor based on the behaviors that load on that factor. The resulting factors and factor loadings are shown in Appendix Tables 14-18. We then estimate the effect of the two programs on each factor with equation (2).

Appendix Table 17 shows results for the reading factor analysis variables. Column 3 shows that full-cost program teachers are more likely to be active throughout the classroom and reduced-cost program teachers are somewhat less likely; the difference between the two versions is significant at the 0.10 level. There is a statistically-significant decline in whole-language exercises at the chalkboard (where the teacher covers all the different literacy concepts at once) in the full-cost classrooms (Column 5), and the difference between the two program versions is significant at the 0.05 level as well. There is also an increase in practicing reading Leblango sentences from readers (Column 7) and paragraphs in primers (Column 8); these changes are larger for the full-cost classrooms but the differences are not statistically significant.

The writing factor analysis results for writing in Appendix Table 18 tell a similar story to the results for the individual components. Teachers in the full-cost treatment arm exhibit an increase in being active throughout the classroom (Column 3) that is at the margin of statistical

significance. The difference across treatment groups is itself statistically significant. Both program versions show drops in copying the teacher's text (Column 5); this effect is again significantly larger among full-cost students. The index for practicing Leblango using slates is massively higher among full-cost program students (Column 7). It also increases substantially for reduced-cost students, reflecting the fact that this index loads on multiple underlying variables and can be positive even in the absence of slates. The estimates possibly reflect the reduced-cost teachers attempt to carry out the NULP pedagogical model, but with limited success because they lack a key input (slates).

Appendix D: Details of Mediation Analysis Methods and Results

The mediation results are presented in Appendix Table 19. Sequential *g*-estimation involves three steps. The first step is to estimate the effects of the mediators on the outcome variable. Second, use those estimates to remove the effects of the mediators from the outcome variable, creating a "demediated" outcome. Third, regress the demediated outcome on the treatment indicator to obtain the estimated effect of the treatment on the outcome, net of the changes in the mediators.

The Acharya, Blackwell, and Sen estimator is only applicable to a single binary treatment variable, so we restrict our attention to a pairwise comparison between the full-cost and reduced-cost versions of the NULP. This allows us to explore the mechanisms behind any differences in outcomes between the two program variants.

We present results using the classroom management and pedagogy factors constructed from the classroom observation data as mediators, but the results are nearly identical if we use the raw classroom observation variables instead. We re-center the mediator variables (i.e., the factor indices) relative to the reduced-cost program, by subtracting off the reduced-cost program mean. We then run the following regression to estimate the effect of the mediators on the outcome:

$$y_{isc} = \beta_0 + \beta_1 FullCost_s + \boldsymbol{M}'_{sco}\tau + FullCost_s * \boldsymbol{M}'_{sco}\lambda + \boldsymbol{I}'_{sco}\pi + \boldsymbol{L}'_s\gamma + \eta y_{isc}^{baseline} + \epsilon_{is} \quad \text{(A)}$$

The notation follows equation 1 in the main paper, but also includes a vector of mediator variables, where *o* indexes a specific observation block (10-minute time period) in classroom *c* and school *s*. We allow the effect of the mediators to vary across study arms by including interaction terms, following Acharya, Blackwell, and Sen (2016). We restrict the predictor variables to enter the estimates linearly.

To consistently estimate τ and λ, we need to satisfy a "no intermediate variable bias"

assumption—that there are no variables omitted from our regression that are affected by the treatment and influence the outcome and also correlated with the mediators. While we cannot guarantee that we have accounted for all potential intermediate confounders, we mitigate this possibility in two ways. First, our vector $\boldsymbol{M}'_{iso}$ includes all the factor variables summarizing the classroom observations data. Second, we control for a vector of intermediate variables $\boldsymbol{I}'_{iso}$ that could be confounders: fixed effects for the block of the classroom observation, the round of the visit, the day of the week, the enumerator who conducted the visit, and a control for the total number of observation blocks for a given classroom observation. While our classroom observation data is extremely rich, making the "no intermediate variable bias" assumption plausible, we cannot rule out all potential violations.[2]

After estimating (A), we then construct a de-mediated value of *y* by subtracting two terms from the raw outcome (y_{isc}): 1) the product of the mediators and the estimated coefficient ($\boldsymbol{M}'_{sco}\hat{\tau}$), and 2) the product of the treatment indicator, the mediators, and the estimated interaction coefficient ($FullCost_s * \boldsymbol{M}'_{sco}\hat{\lambda}$). This yields the following expression:

$$y_{isc}^{demediated} = y_{isc} - \boldsymbol{M}'_{sco}\hat{\tau} - FullCost_s * \boldsymbol{M}'_{sco}\hat{\lambda} \qquad \text{(B)}$$

The result, $y_{isc}^{demediated}$, can be interpreted as the outcome variable purged of the effects of changes

[2] Because we use the factor analysis indices as our mediator variables instead of the raw classroom observation variables, it is conceivable that some of the raw variables could be intermediate confounders. However, if we instead use the full set of raw variables, our results barely change. A potentially important omitted mediator is teacher attendance, which was not collected during the classroom observation visits. Head teacher surveys on teacher attendance suggest no differences across study arms; however, we cannot rule out the possibility of important complementarities of program inputs with teacher attendance in treatment schools.

in the mediator variables. We then can estimate a modified equation 1, regressing the de-mediated value of *y* on the treatment indicator and our baseline controls (recall we are not using control-group observations in these analyses):

$$y_{isc}^{demediated} = \beta_0 + \beta_1 FullCost_s + \boldsymbol{L}_s'\gamma + \eta y_{is}^{baseline} + \epsilon_{is} \qquad \text{(C)}$$

Acharya, Blackwell, and Sen (2016) show that under the assumption of no intermediate variable bias, equation C estimates the *average controlled direct effect*. In our case, this is the difference in test scores between the full-cost treatment and the reduced-cost treatment, under the counterfactual hypothesis that all mediators are held at the mean value in the reduced-cost study arm. This allows us to measure what proportion of the treatment effect can be explained through changes in the mediators we measure through the classroom observations. Specifically, we compare this estimate to the main treatment effect estimates from equation (1) to assess the share of the change in test scores driven by changes in our measured mediators.

Appendix E: Details of Machine Learning Methods and Results

We use two machine-learning methods: KRLS and the LASSO. Due to the computationally-intensive nature of these techniques, we simplify the problem in two ways. First, we focus on the factor analysis indices instead of the raw classroom observation variables. Second, we collapse the data to classroom-level means of both the classroom observation variables, $\bar{\boldsymbol{M}}'_{sc}$, and the exam scores, $\bar{y}_{sc}$. We estimate the following equation:

$$\bar{y}_{sc} = f(\bar{\boldsymbol{M}}'_{sc}) + \epsilon_{sc} \qquad \text{(D)}$$

Here $\bar{y}_{sc}$ is the average endline exam score in classroom *c* in school *s,* $\bar{\boldsymbol{M}}'_{sc}$ is a vector of the average values of the mediators in classroom *c* in school *s*, and $f(\,)$ is a flexible function of the classroom-average mediators.

We approximate $f(\,)$ by including all a third-degree polynomial in each mediator and all interactions between mediators up to third order. The LASSO considers only the predictors we give it directly, and thus is more susceptible to misspecification bias than KRLS (Hainmueller and Hazlett, 2014). Because KRLS explores interactions and higher-order terms automatically, it is able to find useful predictors that we were unable to give the LASSO due to computation time constraints, such as four-way interactions and quartic terms. The results of using machine learning to predict test scores from the classroom observation factors are presented in Panel A of Appendix Table 20.

A potential concern with these estimates is overfitting: it is possible these R-squared values reflect strong predictive power within our sample that would not actually generalize to other datasets. To assess the potential for overfitting, we focus on the KRLS estimates in order to reduce the computing time needed to generate the results. The Townsend (2018) implementation of the LASSO uses the coordinate descent algorithm, which degrades in performance quickly for cases

like ours where there are far more predictors than observations (Friedman et al. 2010). For our data, each run of KRLS takes less than one minute while each LASSO run takes more than an hour; this renders direct testing of overfitting in the LASSO relatively impractical. Fortunately KRLS achieves similar predictive power to the LASSO for reading scores, and much higher predictive power for writing scores, so we interpret our results here as informative as to the extent of overfitting for both techniques.

The KRLS estimator is designed to mitigate overfitting by using leave-one-out cross-validation. If there are K observations it fits the model K times, in each instance leaving out one observation and computing the error in predicting the outcome for that observation. It then selects the functional form that minimizes the sum of the squared leave-one-out errors; the method thus provides a high degree of out-of-sample fit, and the estimated R-squared converges to the true value asymptotically. Overfitting can still occur, however; in small samples (N<100), Hainmuller and Hazlett show that their estimated R-squared may be biased upward.

As a check on the potential for overfitting, we apply the KRLS estimator to random noise. If the estimator yields low R-squared values when applied to noise, then we can infer that it is finding real predictive power in our mediators. To do this test, we replace the real mediators with random numbers, using the same number of random variables as we have mediators in the real data (21 variables). We then use the random numbers as "mediators" to see how well KRLS can use them to predict the outcome; we repeat the process 1000 times and report the average R-squared across all 1000 iterations in Panel B of Appendix Table 20. Replacing the actual data with random noise yields median R-squared values of 0.016 for reading and 0.035 for writing, suggesting that any upward bias in our estimates of the predictive power of the mediators is minimal.

An alternative way to assess the empirical importance of overfitting is to use split-sample cross-validation. We randomly split the sample in half, estimating the model on one-half and assessing the fit (as measured by the R-squared) on the other. The drawback is that this requires estimating the model using a very small sample. We only have 72 classrooms with data on all the mediators, so a 50% random test sample contains just 36 observations. This is likely to be problematic for accurately assessing model fit: Harrell (2015) recommends that test samples have at least 100 observations. We repeat the split-sample approach 1000 times and report the average out-of-sample R-squared across all 1000 iterations in Panel B of Appendix Table 20. Constructing predicted values from the KRLS estimates and using them to predict out-of-sample outcomes gives a mean R-squared of 0.01 for reading and 0.05 for writing. The split-sample results suggest that KRLS could be over-fitting in the full sample, but we believe the small sample sizes involved (just 36 observations) could be driving the low predictive power attained in these out-of-sample checks.

References

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

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

3. Gelman, Andrew, & John Carlin. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651.

4. Glewwe, Paul, Philip Ross, & Bruce Wydick. (2018). Developing Hope among Impoverished Children: Using Child Self-Portraits to Measure Poverty Program Impacts. *Journal of Human Resources*, 53(2), 330–355.

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

6. Harrell Jr., Frank. (2015). Model Validation. In *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis* (pp. 109–116). Switzerland: Springer.

7. Ioannidis, John, T. D. Stanley, & Hristos Doucouliagos. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605), F236–F265.

8. McKenzie, David, & Owen Ozier. (2019). Why ex-post power using estimated effect sizes is bad, but an ex-post MDE is not. *World Bank Development Impact Blog*, May 16, 2019.

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

Appendix Figure 1

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 1

NULP Components and Marginal Costs by Study Arm

	Full-cost program		Reduced-cost program	
	Amount	Cost per Student	Amount	Cost per Student
Pedagogy				
Mother-Tongue-First Instruction	Yes		Yes	
NULP Instructional Model	Yes		Yes	
Books				
Leblango Primers	3/student (1/term)	$0.91	3/student (1/term)	$0.91
Leblango Readers	3/student (1/term)	$0.91	3/student (1/term)	$0.91
Leblango Alphabet	1 per classroom	$0.03		$0.03
Leblango Teacher's	1 per classroom	$0.12	1 per classroom	$0.12
English Primers	3/student (1/term)	$0.91	3/student (1/term)	$0.91
English Teacher's	1 per classroom	$0.12	1 per classroom	$0.12
Materials				
Slates	1 per student	$1.16		$0.00
Wall Clocks	1 per classroom	$0.13		$0.00
Training and Support for				
Literacy Methods Training	4X/year, residential, run by MT staff	$8.82	4X/year, non-residential, run by	$3.51
Saturday in-service training workshops	2X/term, non-residential, run by MT	$3.21	2X/term, non-residential, run by	$0.62
Classroom support supervision	3X/term by MT staff, 2X/term by CCTs	$1.69	2X/term from CCTs	$0.00
Other				
Parent Meetings	1X/term	$1.86		
Take a Book Home Activity	1X/year			
Total Cost		**$19.88**		**$7.14**

Notes: This table shows the components of each version of the NULP intervention and their marginal costs. The costs of developing the intervention and materials are not included as those are one-off costs that will not be repeated in the future. Monetary costs are drawn from a detailed expense workbook shared by Mango Tree, except for the cost of wall clocks, which we estimate from local markets. We also include time costs for teachers (estimated from survey data at $5.74/day) in the Training and Support for Teachers category. Time costs are only counted for days on which the person would not otherwise be working.

Appendix Table 2

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

	Full-cost	Reduced-cost	Other Programs in Arancibia, Popova, and Evans Sample	
			Mean	SD
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		
Targeting by geography	1	1	0.50	0.51
Targeting by subject	0	0	0.27	0.46
Targeting by grade	1	1	0.77	0.43
Targeting by years of experience	0	0	0.05	0.21
Targeting by skill gaps	0	0	0.00	0.00
Targeting by contract teachers	0	0	0.14	0.35
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	0.00	0.00
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	0.11	0.32
Did the program provide storybooks?	1	1	0.21	0.42
Did the program provide computers?	0	0	0.16	0.37
Did the program provide teacher manuals?	1	1	0.53	0.51
Did the program provide lesson plans/videos?	1	1	0.32	0.48
Did the program provide scripted lessons?	1	1	0.00	0.00
Did the program provide craft materials?	0	0	0.16	0.37
Did the program provide other reading materials - flashcards, word banks, reading pamphlets or similar?	1	1	0.26	0.45
Did the program provide software?	0	0	0.30	0.47
How many teachers received training under this program each year?	24	28	706.25	1739.51
How many schools is the program being implemented in (at the time of the evaluation)?	12	14	61.26	43.89
How many years has the program been running (at the time of the evaluation)?	2	2	3.07	3.25
In the last year what percentage of the teachers who began the training dropped out before the end?	0	0	0.21	0.29
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	0.92	0.29
Did the training involve discussion?	1	1	0.58	0.51
Did the training involve lesson enactment?	1	1	0.50	0.52
Did the training involve materials development?	0	0	0.33	0.49
Did the training involve training on how to conduct diagnostics?	1	1	0.23	0.44
Did the training involve lesson planning?	1	1	0.53	0.51
Did the training involve the use of scripted lessons?	1	1	0.25	0.45
Is it a cascade training model (i.e. one where program trainers trainers who then train teachers)?	0	1	0.50	0.51
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	0.91	0.29
During this period, what is the total hours of teacher training they receive?	120	120	59.17	44.23
During this period, how many hours of lectures do they receive?	60	60	26.45	23.86
During this period, how many hours do they spend practicing with students?	0	0	7.39	8.56
During this period, how many hours do they spend practicing with other teachers?	60	60	45.56	35.73
Over how many weeks?	40	40	9.23	12.18
Where does this part of the training take place?	2	2		
How many teachers are in each training session?	24	26	27.85	8.57
How many in-school follow-up support visits do teachers receive after the initial training (if any)?	9	6	5.77	10.05
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	1.64	4.62
Over how many months?	9	9	11.83	8.71
Africa dummy	1	1	0.21	0.41
Interviewed	1	1	0.42	0.50

Notes: Data comes from Arancibia, Popova, and Evans (2016). The means and standard deviations are included only for dummy variables and numbers; we omit the statistics for the categorical numerical fields as they cannot be meaningfully interpreted. † The full-cost NULP was coded as Primary or secondary teachers on this indicator (1), while the reduced-cost version was coded as Local government official (4).

Appendix Table 3

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	0.000	0.039	-0.085	-0.000	-0.130	-0.045
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	0.010	-0.008	-0.024	-0.039	-0.022	-0.036
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
Number of students	600	600	700	477	485	519	123	115	181
Number of schools	12	12	14	12	12	14	12	12	14

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 randomization inference 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 4

Predictors of Attrition by Study Arm

	(1)	(2)	(3)	(4)
	Outcome: Attritted			
	Control	Full-cost	Reduced-cost	All
Female	0.000	0.024	0.055	0.000
	(0.036)	(0.030)	(0.033)	(0.035)
Female*(Full-cost Program)				0.055
				(0.047)
Female*(Reduced-cost Program)				0.024
				(0.046)
Age	0.004	0.011	0.022	0.004
	(0.014)	(0.012)	(0.015)	(0.014)
Age*(Full-cost Program)				0.018
				(0.020)
Age*(Reduced-cost Program)				0.008
				(0.018)
PCA Leblango EGRA Score Index	0.007	0.049	-0.031***	0.007
	(0.027)	(0.044)	(0.008)	(0.026)
PCA Leblango EGRA Score Index*(Full-cost Program)				-0.037
				(0.027)
PCA Leblango EGRA Score Index*(Reduced-cost Program)				0.042
				(0.050)
PCA Writing EGRA Score Index	-0.357***	-0.323*	-0.258	-0.357***
	(0.092)	(0.180)	(0.159)	(0.089)
PCA Writing EGRA Score Index*(Full-cost Program)				0.099
				(0.178)
PCA Writing EGRA Score Index*(Reduced-cost Program)				0.034
				(0.198)
Full-cost Program				-0.172
				(0.161)
Reduced-cost Program				-0.015
				(0.143)
Number of students	589	661	588	1,838
Number of schools	12	14	12	38
Adjusted R-squared	0.011	0.001	0.010	0.011

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.

Appendix Table 5

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 Recognition	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**
S.E.	(0.127)	(0.163)	(0.129)	(0.091)	(0.097)	(0.121)	(0.110)
R.I. p-value	[0.004]	[0.004]	[0.007]	[0.004]	[0.135]	[0.015]	[0.028]
q-value	--	{0.024}	{0.028}	{0.024}	{0.231}	{0.045}	{0.067}
Reduced-cost program	0.110	0.418	0.064	-0.012	0.021	0.058	0.034
S.E.	(0.102)	(0.181)	(0.096)	(0.074)	(0.069)	(0.081)	(0.084)
R.I. p-value	[0.367]	[0.104]	[0.513]	[0.862]	[0.790]	[0.516]	[0.730]
q-value	--	{0.208}	{0.688}	{0.862}	{0.862}	{0.688}	{0.862}
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 treatment effects	0.544***	0.624**	0.585***	0.393***	0.213	0.426**	0.415**
S.E.	(0.124)	(0.159)	(0.127)	(0.092)	(0.093)	(0.115)	(0.120)
R.I. p-value	[0.006]	[0.017]	[0.007]	[0.001]	[0.127]	[0.012]	[0.031]
q-value	--	{0.025}	{0.021}	{0.006}	{0.127}	{0.024}	{0.037}
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 6

Wild Cluster Bootstrap p-values for Program Impacts on Leblango Early Grade Reading Assessment 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 Recognition	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Full-cost program	0.638***	1.014***	0.647***	0.374**	0.215	0.476**	0.445***
S.E.	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.128)	(0.113)
Wild Cluster Bootstrap p-value	[0.006]	[0.001]	[0.002]	[0.047]	[0.175]	[0.037]	[0.009]
Reduced-cost program	0.129	0.407*	0.076	-0.002	0.031	0.071	0.045
S.E.	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
Wild Cluster Bootstrap p-value	[0.133]	[0.046]	[0.243]	[0.712]	[0.475]	[0.309]	[0.272]
Number of students	1460	1476	1481	1474	1471	1467	1481
Number of schools	38	38	38	38	38	38	38
Adjusted R-squared	0.149	0.219	0.103	0.066	0.075	0.074	0.058
Difference between treatment effects	0.509*	0.607*	0.570**	0.376**	0.184	0.405*	0.400**
S.E.	(0.127)	(0.159)	(0.128)	(0.092)	(0.093)	(0.117)	(0.120)
Wild Cluster Bootstrap p-value	[0.054]	[0.083]	[0.011]	[0.042]	[0.301]	[0.083]	[0.015]
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 7

Lee Bounds for Program Impacts on Early Grade Reading Assessment Scores
(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Leblango EGRA Score Index†	(2) Letter Name Knowledge	(3) Initial Sound Recognition	(4) Familiar Word Recognition	(5) Invented Word Recognition	(6) Oral Reading Fluency	(7) Reading Comprehension
Full-cost program							
Lee Upper Bound	0.642***	1.045***	0.659***	0.386***	0.223**	0.480***	0.460***
S.E.	(0.136)	(0.169)	(0.132)	(0.098)	(0.104)	(0.129)	(0.115)
Lee Lower Bound	0.558***	0.955***	0.602***	0.194**	0.100	0.341***	0.350***
S.E.	(0.115)	(0.167)	(0.124)	(0.074)	(0.072)	(0.108)	(0.098)
Reduced-cost program							
Lee Upper Bound	0.282**	0.590***	0.304***	0.138*	0.139**	0.200**	0.206**
S.E.	(0.105)	(0.174)	(0.096)	(0.074)	(0.067)	(0.088)	(0.083)
Lee Lower Bound	0.108	0.364**	0.047	-0.017	0.019	0.062	0.007
S.E.	(0.104)	(0.178)	(0.096)	(0.078)	(0.070)	(0.085)	(0.088)
Raw (unadjusted) values§							
Control group mean	0.144	5.973	0.616	0.334	0.358	0.611	0.216
Control group SD	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Leblango EGRA Score Index is constructed by weighting each of the 6 test modules (columns 2 through 7) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 8

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)	English Name (Given Name)	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Full-cost program	0.399	1.015***	1.230***	0.147	0.442	0.152	0.128	0.377	0.221	0.139
S.E.	(0.186)	(0.116)	(0.148)	(0.178)	(0.207)	(0.156)	(0.178)	(0.210)	(0.173)	(0.150)
R.I. p-value	[0.168]	[0.001]	[0.001]	[0.626]	[0.173]	[0.539]	[0.663]	[0.251]	[0.385]	[0.558]
q-value	--	{0.009}	{0.009}	{0.663}	{0.283}	{0.628}	{0.663}	{0.377}	{0.495}	{0.628}
Reduced-cost program	-0.232	0.437**	0.393*	-0.288	-0.317	-0.313***	-0.308*	-0.334*	-0.253	-0.330***
S.E.	(0.163)	(0.127)	(0.152)	(0.150)	(0.178)	(0.134)	(0.151)	(0.179)	(0.156)	(0.129)
R.I. p-value	[0.407]	[0.020]	[0.061]	[0.153]	[0.155]	[0.006]	[0.096]	[0.096]	[0.297]	[0.007]
q-value	--	{0.072}	{0.183}	{0.279}	{0.279}	{0.032}	{0.216}	{0.216}	{0.411}	{0.032}
Number of students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.265	0.193	0.217	0.161	0.304	0.177	0.165	0.300	0.164	0.171
Difference between treatment	0.631***	0.577**	0.837***	0.435***	0.758***	0.465***	0.436***	0.711***	0.474***	0.469***
S.E.	(0.149)	(0.136)	(0.156)	(0.151)	(0.173)	(0.118)	(0.150)	(0.175)	(0.151)	(0.115)
R.I. p-value	[0.000]	[0.014]	[0.001]	[0.005]	[0.000]	[0.003]	[0.006]	[0.001]	[0.005]	[0.003]
q-value	--	{0.014}	{0.003}	{0.006}	{0.000}	{0.005}	{0.007}	{0.003}	{0.006}	{0.005}
Raw (unadjusted) values§										
Control group mean	0.482	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control group SD	1.000	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 9

Wild Cluster Bootstrap p-values for Program Impacts on Writing Test Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA	Name-Writing		Story-Writing						
	Writing Score Index†	African (Family) Name	English (Given) Name	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Full-cost program	0.449**	0.922***	1.312***	0.163	0.441	0.152	0.175	0.383	0.221	0.139
S.E.	(0.144)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
Wild Cluster Bootstrap p-value	[0.032]	[0.000]	[0.000]	[0.525]	[0.207]	[0.531]	[0.431]	[0.280]	[0.433]	[0.562]
Reduced-cost program	-0.159	0.435***	0.450***	-0.274	-0.316	-0.313	-0.262	-0.330	-0.253	-0.33
S.E.	(0.122)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
Wild Cluster Bootstrap p-value	[0.623]	[0.005]	[0.007]	[0.344]	[0.361]	[0.214]	[0.213]	[0.294]	[0.537]	[0.156]
Number of students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.352	0.240	0.236	0.174	0.304	0.177	0.200	0.302	0.164	0.171
Difference between treatment effects	0.608***	0.487**	0.861***	0.436**	0.757***	0.465***	0.437**	0.713***	0.474***	0.469***
S.E.	(0.128)	(0.135)	(0.154)	(0.148)	(0.173)	(0.118)	(0.139)	(0.174)	(0.151)	(0.115)
Wild Cluster Bootstrap p-value	[0.001]	[0.008]	[0.000]	[0.020]	[0.002]	[0.003]	[0.016]	[0.002]	[0.008]	[0.003]
Raw (unadjusted) values§										
Control group mean	0.482	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control group SD	1.000	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Presentation (column 10), which was not one of the marked categories at baseline; missing values of control variables are dummied out. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 10

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)	English Name (Given Name)	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**
S.E.	(0.108)	(0.117)	(0.150)	(0.109)	(0.129)	(0.091)	(0.114)	(0.130)	(0.110)	(0.088)
R.I. p-value	[0.006]	[0.001]	[0.001]	[0.056]	[0.003]	[0.028]	[0.082]	[0.003]	[0.020]	[0.030]
q-value	--	{0.009}	{0.009}	{0.084}	{0.014}	{0.049}	{0.114}	{0.014}	{0.047}	{0.049}
Reduced-cost program	-0.004	0.473**	0.527***	-0.093	-0.079	-0.130**	-0.107	-0.093	-0.050	-0.155**
S.E.	(0.076)	(0.125)	(0.149)	(0.078)	(0.088)	(0.060)	(0.078)	(0.085)	(0.082)	(0.060)
R.I. p-value	[0.960]	[0.011]	[0.004]	[0.309]	[0.328]	[0.024]	[0.197]	[0.217]	[0.608]	[0.021]
q-value	--	{0.033}	{0.014}	{0.347}	{0.347}	{0.048}	{0.253}	{0.260}	{0.608}	{0.047}
Number of students	1262	1336	1263	1361	1361	1360	1360	1361	1361	1361
Number of schools	35	35	35	35	35	35	35	35	35	35
Adjusted R-squared	0.315	0.234	0.241	0.153	0.319	0.165	0.151	0.302	0.146	0.158
Difference between treatment	0.618***	0.460**	0.837***	0.464***	0.780***	0.480***	0.458***	0.731***	0.485***	0.484***
S.E.	(0.117)	(0.144)	(0.162)	(0.130)	(0.146)	(0.091)	(0.127)	(0.147)	(0.130)	(0.090)
R.I. p-value	[0.004]	[0.040]	[0.004]	[0.001]	[0.000]	[0.000]	[0.008]	[0.000]	[0.002]	[0.000]
q-value	--	{0.040}	{0.005}	{0.002}	{0.000}	{0.000}	{0.009}	{0.000}	{0.003}	{0.000}
Raw (unadjusted) values§										
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	0.310	0.190	0.298

Notes: Longitudinal sample includes 1,361 students from 35 schools who were tested at baseline as well as endline and are not from the stratification cell where one school conducted the writing test in English. 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. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 11

Lee Bounds for Program Impacts on Writing Test Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	African Name (Surname)	English Name (Given Name)	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Full-cost program										
Lee Upper Bound	0.512***	0.959***	1.448***	0.170	0.458**	0.158	0.190	0.395*	0.232	0.150
S.E.	(0.149)	(0.105)	(0.145)	(0.173)	(0.207)	(0.158)	(0.148)	(0.206)	(0.173)	(0.149)
Lee Lower Bound	0.305**	0.901***	1.186***	0.111	0.409*	0.107	0.102	0.340*	0.170	0.098
S.E.	(0.124)	(0.104)	(0.143)	(0.162)	(0.202)	(0.154)	(0.136)	(0.200)	(0.166)	(0.144)
Reduced-cost program										
Lee Upper Bound	-0.094	0.605***	0.546***	-0.103	-0.120	-0.148	-0.117	-0.135	-0.059	-0.175*
S.E.	(0.099)	(0.114)	(0.142)	(0.107)	(0.092)	(0.095)	(0.091)	(0.089)	(0.103)	(0.100)
Lee Lower Bound	-0.183	0.375***	0.420***	-0.307*	-0.348*	-0.344**	-0.285**	-0.361*	-0.276*	-0.362**
S.E.	(0.124)	(0.121)	(0.151)	(0.152)	(0.183)	(0.140)	(0.129)	(0.185)	(0.163)	(0.135)
Raw (unadjusted) values§										
Control group mean	0.482	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control group SD	1.000	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students from 38 schools who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001. Benjamini and Yekutieli (2001) q-values, which adjust the p-values to control the false discovery rate, in braces.

† PCA Writing Score Index is constructed by weighting each of the 9 test modules (columns 2 through 10) using the first principal component of the 2013 endline control-group data as in Black and Smith (2006), normalized by dividing by the endline control-group standard deviation. § Control Group Mean and SD are the raw (unstandardized) means and SDs computed using the endline data for control-group observations in the estimation sample.

Appendix Table 12

Productivity of Time on Task

	(1) Full-cost program	(2) Reduced-cost program	(3) Control
Total literacy class time in P1			
# of terms	3	3	3
Instruction weeks per term	12	12	12
Classes per week	10	10	10
Minutes Per class	30	30	30
Total literacy hours in P1	180	180	180
Reading			
Share of time spent on reading	0.379	0.370	0.318
Total hours spent on reading	68.2	66.6	57.2
Reading gain in P1	0.786	0.277	0.148
Reading gain per hour	0.012	0.004	0.003
Writing			
Share of time spent on writing	0.209	0.242	0.241
Total hours spent on reading	37.6	43.6	43.4
Writing gain in P1	0.917	0.309	0.468
Writing gain per hour	0.024	0.007	0.011

Notes: This table combines information on time use from Table 5 with the estimated gains in reading and writing by study arm from Tables 2 and 3 to estimate the productivity of each minute of class time during first grade.

Appendix Table 13

Classroom Observations: Elements of Focus

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Element of Focus During Reading				Element of Focus During Writing					Copying Text from	Writing Own Text
	Sounds	Letters	Words	Sentences	Pictures	Letters	Words	Sentences	Name		
Full-cost program	0.106**	0.048	0.054	0.094*	0.107	-0.085	0.017	-0.011	0.136	-0.165**	0.241**
S.E.	(0.020)	(0.028)	(0.034)	(0.036)	(0.052)	(0.044)	(0.051)	(0.052)	(0.055)	(0.046)	(0.054)
R.I. p-value	[0.011]	[0.301]	[0.333]	[0.050]	[0.191]	[0.213]	[0.802]	[0.872]	[0.111]	[0.021]	[0.011]
q-value	{0.083}	{0.529}	{0.529}	{0.177}	{0.357}	{0.357}	{0.856}	{0.872}	{0.256}	{0.126}	{0.126}
Reduced-cost program	0.075**	0.082	0.024	0.017	0.110*	0.040	0.126*	-0.074	0.099**	-0.036	0.071
S.E.	(0.021)	(0.031)	(0.033)	(0.043)	(0.049)	(0.042)	(0.051)	(0.041)	(0.034)	(0.048)	(0.046)
R.I. p-value	[0.015]	[0.113]	[0.509]	[0.826]	[0.072]	[0.542]	[0.065]	[0.199]	[0.027]	[0.549]	[0.276]
q-value	{0.090}	{0.308}	{0.694}	{0.826}	{0.205}	{0.659}	{0.205}	{0.357}	{0.135}	{0.659}	{0.394}
Number of lessons	398	398	398	398	326	326	326	326	326	326	326
Number of schools	38	38	38	38	38	38	38	38	38	38	38
Adjusted R-squared	0.099	0.016	0.101	0.075	0.091	0.115	0.186	0.211	0.294	0.202	0.282
Difference between treatment effects	0.031	-0.034	0.030	0.077**	-0.003	-0.125*	-0.108	0.063	0.037	-0.128**	0.169**
S.E.	(0.018)	(0.027)	(0.034)	(0.028)	(0.043)	(0.038)	(0.043)	(0.043)	(0.048)	(0.044)	(0.046)
R.I. p-value	[0.280]	[0.340]	[0.502]	[0.022]	[0.949]	[0.054]	[0.115]	[0.266]	[0.573]	[0.032]	[0.012]
q-value	{0.600}	{0.637}	{0.685}	{0.240}	{0.963}	{0.162}	{0.246}	{0.499}	{0.811}	{0.120}	{0.060}
Control group mean	0.046	0.161	0.622	0.320	0.181	0.194	0.326	0.160	0.094	0.368	0.142
Control group SD	0.132	0.237	0.310	0.320	0.241	0.285	0.274	0.251	0.220	0.304	0.209

Notes: Sample is 398 lessons in which students do any reading and 326 lessons in which students do any writing, based on 440 lesson observations for 38 schools. All regressions control for indicators for stratification cell, the round of the observations the enumerator, and the day of the week, as well as the average value of the observation period (1, 2, or 3) for the lesson, and are weighted by the share of time spent on reading (columns 1-2) or writing (columns 3-7) during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01.

Appendix Table 14

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. Classroom management variables are measured in general for each observation window and are not specific to reading or writing activities, so we estimate the factors on the pooled sample of all lessons. 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 15

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 16

Factor Loadings for Writing Pedagogy Indices

	(1)	(2)	(3)	(4)	(5)
	Pictures, Words, and Stories	Copying Teacher's Text	Pictures and Letters on Paper, High-Energy	Leblango Practice on Slates	Leblango Sentences and Handwriting
Students are Writing:					
Pictures	0.15	-0.04	0.12	0.11	-0.14
Letters	-0.50	0.04	0.11	0.15	-0.08
Words	0.10	0.11	-0.07	0.04	-0.04
Sentences	0.04	0.05	0.03	-0.02	0.34
Their Names	0.06	0.00	0.00	0.24	0.07
Air Writing	-0.22	-0.13	-0.05	0.00	0.04
Handwriting Practice	-0.01	0.02	0.03	0.15	0.26
Copying Teacher's Text from Board	0.05	0.44	0.03	0.09	-0.04
Writing Own Text	0.12	-0.34	0.07	0.08	-0.07
On Slate	0.01	0.00	-0.11	0.31	-0.03
On Paper	0.06	0.06	0.39	-0.11	0.04
On Board	0.00	-0.02	-0.22	-0.11	-0.01
Percent of Students Participating	-0.01	0.01	0.14	0.08	-0.12
Share in Leblango	-0.05	-0.06	0.01	0.18	0.11
Share of Variance Explained	0.46	0.31	0.16	0.21	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 17

Effects on Pedagogy and Classroom Management Factor Indices for Reading Classes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Classroom Management					Pedagogy		
	Keeps Students Focused	Solid Lesson Plan	Active Throughout Classroom	Sounds and Letters Only	Whole Language On Board	Basic Elements in Breakout Sessions	Leblango Sentences in Reader	Paragraphs in Primer
Full-cost program	-0.120	0.040	0.045	0.111	-0.189**	0.026	0.254**	0.241**
S.E.	(0.079)	(0.048)	(0.057)	(0.046)	(0.064)	(0.065)	(0.050)	(0.050)
R.I. p-value	[0.223]	[0.635]	[0.456]	[0.129]	[0.046]	[0.817]	[0.010]	[0.017]
Reduced-cost program	-0.114	0.035	-0.034	0.147	-0.007	0.019	0.151**	0.159**
S.E.	(0.082)	(0.048)	(0.048)	(0.058)	(0.048)	(0.073)	(0.050)	(0.048)
R.I. p-value	[0.324]	[0.595]	[0.564]	[0.133]	[0.912]	[0.853]	[0.036]	[0.030]
Number of observation periods	398	398	398	398	398	398	398	398
Adjusted R-squared	0.156	0.161	0.372	0.033	0.133	0.180	0.229	0.114
Difference between treatment effects	-0.006	0.005	0.079*	-0.035	-0.181**	0.007	0.102	0.082
S.E.	(0.097)	(0.045)	(0.035)	(0.041)	(0.063)	(0.075)	(0.048)	(0.046)
R.I. p-value	[0.965]	[0.927]	[0.083]	[0.416]	[0.037]	[0.935]	[0.168]	[0.265]
Control group mean	0.141	0.008	-0.026	-0.120	-0.185	-0.458	-0.253	-0.426
Control group SD	0.487	0.477	0.443	0.357	0.528	0.610	0.355	0.404

Notes: Sample is 398 lessons in which students do any reading and 326 lessons in which students do any writing, based on 440 lesson observations for 38 schools. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Appendix Table 18

Effects on Pedagogy and Classroom Management Factor Indices for Writing Classes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Classroom Management					Pedagogy		
	Keeps Students Focused	Solid Lesson Plan	Active Throughout Classroom	Pictures, Words, and Stories	Copying Teacher's Text	Pictures and Letters on Paper, High-Energy	Leblango Practice on Slates	Leblango Sentences and Handwriting
Full-cost program	-0.152	0.093	0.152	0.168*	-0.349***	-0.033	0.398**	-0.012
S.E.	(0.083)	(0.059)	(0.064)	(0.066)	(0.060)	(0.073)	(0.102)	(0.056)
R.I. p-value	[0.145]	[0.430]	[0.101]	[0.088]	[0.002]	[0.733]	[0.027]	[0.843]
Reduced-cost program	-0.180*	0.125*	0.039	-0.017	-0.100	0.119	0.254***	-0.038
S.E.	(0.079)	(0.057)	(0.053)	(0.061)	(0.068)	(0.067)	(0.073)	(0.052)
R.I. p-value	[0.088]	[0.053]	[0.553]	[0.856]	[0.288]	[0.122]	[0.000]	[0.649]
Number of observation periods	326	326	326	326	326	326	326	326
Adjusted R-squared	0.156	0.302	0.315	0.177	0.247	0.209	0.215	0.283
Difference between treatment	0.028	-0.032	0.113**	0.185*	-0.249***	-0.152	0.144	0.026
S.E.	(0.087)	(0.077)	(0.044)	(0.060)	(0.056)	(0.083)	(0.089)	(0.047)
R.I. p-value	[0.830]	[0.793]	[0.018]	[0.062]	[0.004]	[0.148]	[0.165]	[0.712]
Control group mean	0.089	-0.010	0.002	-0.099	-0.061	-0.639	-0.775	0.044
Control group SD	0.532	0.502	0.430	0.459	0.379	0.417	0.368	0.305

Notes: Sample is 398 lessons in which students do any reading and 326 lessons in which students do any writing, based on 440 lesson observations for 38 schools. All regressions control for indicators for stratification cell, the round of the observations, the period of the observation block (1, 2, or 3), the enumerator, and the day of the week, and are weighted by the share of time spent on reading during the observation window. Control Group Mean and SD are computed using the pooled data for the control group across all three rounds of classroom observations. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.05, ** p<0.01, *** p<0.001.

Appendix Table 19

Mediation Analysis

	(1) Letter Name Knowledge	(2) PCA Leblango EGRA Score Index	(3) PCA Writing Score Index
Demediated Treatment Effect			
Difference between full-cost and reduced-cost programs	0.681***	0.598***	0.645***
S.E.	(0.127)	(0.095)	(0.101)
R.I. p-value	[0.002]	[0.000]	[0.000]
Adjusted R-squared	0.232	0.159	0.331
Number of observations	15,516	15,311	14,559
Share of treatment effect explained by mediators	0.011	0.020	0.037
Raw (unadjusted) values[§]			
Reduced-cost program mean	11.346	0.31	-0.054
Reduced-cost program SD	13.861	1.072	0.639

Notes: Sample is the combination of each student with all classroom observation windows for that student's class; re-estimating our main regressions on this modified sample yields similar treatment effects and confidence intervals to the main sample. The analyses in this table are restricted to data from the two treatment arms. We estimate the demediated treatment effect using the sequential *g* estimator of Acharya et al. (2016), by removing the effect of the treatment on the mediators from the outcome and then regressing the demediated outcome on the treatment indicator. We reduce the dimensionality of the predictor variables by using the factor analysis indices rather than the raw variables. Reduced-Cost Program Mean and SD are computed using the endline data for the reduced-cost group alone. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Randomization inference p-values, clustered by school and stratified by stratification cell, in brackets; * p<0.1, ** p<0.05, *** p<0.01.

Appendix Table 20

Machine Learning Results

Panel A: Predictive Power by Model

Method	Exam	(1) R-Squared
OLS	Reading	0.024
OLS	Writing	0.037
KRLS	Reading	0.182
KRLS	Writing	0.433
LASSO	Reading	0.197
LASSO	Writing	0.059

Panel B: Tests of Overfitting for KRLS

Test	Exam	Mean R-Squared
Random Predictors Instead of Real Variables	Reading	0.016
Random Predictors Instead of Real Variables	Writing	0.035
Split-Sample Out-of-Sample Predictions	Reading	0.012
Split-Sample Out-of-Sample Predictions	Writing	0.047

Notes: The results in Panel A come from collapsing the mediators and exam scores to classroom-level means and then using the mediators to predict the classroom-average exam scores. We scale down the resulting R-squared by the classroom-level fraction of the overall variance of test scores. For the KRLS and LASSO estimates, we provide the algorithm with third-degree polynomials in each mediator and all two-way interactions; for OLS we enter each mediator linearly. We reduce the dimensionality of the predictor variables by using the factor analysis indices rather than the raw variables. The overfitting tests in Panel B create random predictors or do randomized split-sample cross-validation 1000 times. We report the mean R-squared across all 1000 iterations.

Appendix Table 21

Top 10 Most Important Variables for Predicting EGRA Scores in Machine Learning Results, Pooled Sample

	LASSO		KRLS	
Rank	Variable	Coef.	Variable	Coef.
1	(Speaking & Listening - Group Only) X(Writing - Leblango Practice on Slates) X(Reading - Leblango Sentences in Reader)	-0.216	(Writing - Leblango Practice on Slates) X(Teacher - Keeps Students Focused) X(Teacher - Keeps Students Focused)	0.001511
2	(Reading - Leblango Sentences in Reader)	0.146	(Reading - Leblango Sentences in Reader) X(Reading - Whole Language on Board) X(Reading - Whole Language on Board)	0.001222
3	(Writing - Leblango Practice on Slates) X(Teacher - Keeps Students Focused)	-0.146	(Speaking & Listening - Individual, Teacher, & Group) X(Reading - Sounds & Letters Only) X(Reading - Sounds & Letters Only)	0.000968
4	(Writing - Leblango Practice on Slates) X(Teacher - Keeps Students Focused) X(Teacher - Keeps Students Focused)	0.138	(Writing - Pictures & Letters on Paper, High-Energy) X(Writing - Pictures, Words, & Stories) X(Reading - Leblango Sentences in Reader)	0.000731
5	(Writing - Pictures & Letters on Paper, High-Energy) X(Reading - Paragraphs in Primer) X(Teacher - Active Throughout Classroom)	0.126	(Reading - Leblango Sentences in Reader) X(Reading - Whole Language on Board) X(Teacher - Keeps Students Focused)	0.000729
6	(Reading - Leblango Sentences in Reader) X(Reading - Basic Elements in Breakout Sessions) X(Reading - Sounds & Letters Only)	0.119	(Speaking & Listening - Group Only) X(Teacher - Keeps Students Focused) X(Teacher - Keeps Students Focused)	-0.000653
7	(Writing - Leblango Practice on Slates) X(Reading - Basic Elements in Breakout Sessions) X(Pct Time Teaching)	-0.099	(Speaking & Listening - Group Only) X(Writing - Pictures & Letters on Paper, High-Energy) X(Reading - Leblango Sentences in Reader)	-0.000621
8	(Writing - Pictures & Letters on Paper, High-Energy) X(Writing - Copying Teacher's Text) X(Reading - Leblango Sentences in Reader)	-0.091	(Writing - Pictures, Words, & Stories)	0.000596
9	(Teacher - Keeps Students Focused)	-0.089	(Reading - Basic Elements in Breakout Sessions) X(Reading - Whole Language on Board)	0.000564
10	(Reading - Basic Elements in Breakout Sessions)	-0.082	(Reading - Sounds & Letters Only) X(Teacher - Solid Lesson Plan) X(Pct Time Spent on Reading)	-0.000563

Notes: This table presents the ten most important variables selected by the LASSO and KRLS algorithms for predicting EGRA scores. The coefficients are standardized such that their interpretation is the effect of a one-SD change in the predictor on reading scores measured in SDs. We reduce the dimensionality of the predictor variables by using the factor analysis indices rather than the raw variables.

Appendix Table 22

Top 10 Most Important Variables for Predicting Writing Test Scores in Machine Learning Results, Pooled Sample

	LASSO		KRLS	
Rank	Variable	Coef.	Variable	Coef.
1	(Speaking & Listening - Group Only) X(Reading - Whole Language on Board) X(Reading - Sounds & Letters Only)	-0.142	(Speaking & Listening - Group Only) X(Teacher - Keeps Students Focused) X(Teacher - Keeps Students Focused)	-0.001257
2	(Speaking & Listening - Group Only)	-0.064	(Reading - Basic Elements in Breakout Sessions) X(Pct Time Outside Class) X(Pct Time Outside Class)	-0.001208
3	(Speaking & Listening - Group Only) X(Writing - Leblango Practice on Slates) X(Reading - Leblango Sentences in Reader)	-0.056	(Reading - Leblango Sentences in Reader) X(Reading - Whole Language on Board) X(Reading - Whole Language on Board)	0.001119
4	(Reading - Whole Language on Board) X(Reading - Whole Language on Board) X(Reading - Whole Language on Board)	-0.039	(Speaking & Listening - Individual, Teacher, & Group) X(Reading - Sounds & Letters Only) X(Reading - Sounds & Letters Only)	0.001072
5	(Reading - Sounds & Letters Only) X(Teacher - Keeps Students Focused) X(Pct Time Spent on Reading)	0.035	(Writing - Pictures, Words, & Stories)	0.000994
6	(Reading - Sounds & Letters Only) X(Pct Time Spent on Reading)	0.033	(Reading - Sounds & Letters Only) X(Teacher - Active Throughout Classroom)	-0.000827
7	(Writing - Leblango Practice on Slates) X(Teacher - Keeps Students Focused)	-0.025	(Writing - Leblango Practice on Slates) X(Teacher - Keeps Students Focused) X(Teacher - Keeps Students Focused)	0.000826
8	(Reading - Leblango Sentences in Reader) X(Reading - Whole Language on Board) X(Reading - Whole Language on Board)	0.021	(Reading - Whole Language on Board) X(Reading - Whole Language on Board) X(Pct Time Spent on Reading)	-0.000798
9	(Reading - Leblango Sentences in Reader) X(Reading - Basic Elements in Breakout Sessions) X(Pct Time Outside Class)	0.019	(Speaking & Listening - Group Only) X(Reading - Whole Language on Board) X(Pct Time Spent on Reading)	-0.000766
10	(Writing - Pictures & Letters on Paper, High-Energy) X(Pct Time Spent on Reading) X(Pct Time Outside Class)	-0.012	(Writing - Pictures & Letters on Paper, High-Energy) X(Reading - Paragraphs in Primer) X(Pct Time Outside Class)	-0.000736

Notes: This table presents the ten most important variables selected by the LASSO and KRLS algorithms for predicting writing test scores. The coefficients are standardized such that their interpretation is the effect of a one-SD change in the predictor on reading scores measured in SDs. We reduce the dimensionality of the predictor variables by using the factor analysis indices rather than the raw variables.