How to Build a Reader: Evidence from a Scalable Literacy Intervention in Ghana

Erik T. J. Andersen, Simon Graffy,
Jason T. Kerwin, and Monica Lambon-Quayefio*

January 8, 2026

Abstract

Addressing the massive test score gaps between rich and poor countries will require programs that are both high-impact and scalable. We use an RCT in low-fee private schools in Ghana to study a program that meets both needs. TFLI increases test scores by 0.5 SDs after just 9 months of intervention. We develop a model in which basic skills constrain the development of advanced skills, which predicts the pattern of effects we observe across early reading capabilities, and makes forecasts about the future impacts of the program as it continues into second grade. Moreover, we show that TFLI's impacts scale roughly linearly with time as compared to a shorter-term, smaller-scale pilot RCT. The program's developers use generative AI to accelerate lesson plan development and adaptation to new settings. An observational pilot test of this adaptation to Uganda yields comparable results to our RCT.

JEL Codes: I21, I25, O15

Keywords: Economics of Education, The Learning Crisis, Structured Pedagogy, Differentiated Instruction, Human Capital, Generative AI, Development Economics

*Andersen: Department of Economics, University of Washington (eander46@uw.edu); Graffy: Inspiring Teachers (simon.graffy@inspiringteachers.org); Kerwin: Department of Economics, University of Washington (jkerwin@uw.edu; Lambon-Quayefio: University of Ghana and J-PAL. We thank Noam Angrist, Peter Bergman, Alex Eble, Owen Ozier, and Rebecca Thornton, and seminar audiences from GSF, the Gates Foundation Global Education Team and the Prevail Fund for helpful comments and suggestions. The authors used ChatGPT 5.2 Thinking to accelerate the drafting and revision of the paper. This project would not have been possible without the efforts of Dr. Emmanuel K. Larbi Mantey, Regina Mensah, Mark Valentine Aikins, Philip Semaha, Isaac Gaisey, Victor Ayanful, Tom Lewis, Lukasz Labedzki, and the entire Inspiring Teachers staff. Charles Oloa, Chloe Lewis, and Emmanuel Bentsi Parker Jr. provided excellent fieldwork supervision, and we are also thankful for the hard work of our data collection team. We are also indebted to the teachers and students who generously gave their time to take part in the study. Inspiring Teachers received program funding from the IDP Foundation and the Global Schools Forum to run the intervention. The Peter Cundill Foundation, Windfall Trust, Mulago Foundation, Pace Able, The Bisconti Family Foundation, Prevail Fund and others also provide generous broader support that helped make this program possible. Research activities were supported by the Bill and Melinda Gates Foundation through the Abdul Latif Jameel Poverty Action Lab (J-PAL)'s Learning for All Initiative and by the Jacobs Foundation through Innovations for Poverty Action's Partnership for Technology Program. This study was reviewed and approved by the University of Ghana Ethics Committee for the Humanities (protocol # ECH 278/ 23-24), the University of Washington Human Subjects Division (protocol # STUDY00021036), and Heartland IRB (protocol # 062424-573); it is registered with the AEA RCT Registry under registration number AEARCTR-0014627.

The learning gaps between the world's poorest and richest countries are staggering. In 2021, South African fourth-graders scored 2.99 standard deviations (SDs) lower on reading than their counterparts in Singapore (PIRLS 2021). This gap is almost four times as large as the black-white test score gap for fourth graders in the United States (NAEP 2022). Indeed, these gaps are so massive that it is rare to even measure the richest and poorest countries using the same exam: South Africa is the only country from sub-Saharan Africa that participated in the PIRLS, and is one of the richest countries on the continent. A 3-SD gap between the poorest and richest countries is therefore likely to be a lower bound.

Addressing these gaps will require changes that are far more ambitious than the ones that governments have typically tried in the past. The average education program raises reading test scores by just 0.20 SD (Evans and Yuan 2022); we would have to repeat such a program more than 15 times to bring African reading scores up to rich-country levels. Whether these gains are cost-effective in dollar terms matters very little, as it is logistically impossible to actually run one of these programs over a dozen times.

We present the initial results from a promising approach to this challenge. Our data comes from a randomized trial we conducted with the education non-profit Inspiring Teachers, in which we evaluated their program, Tools for Foundational Learning Improvement (TFLI). TFLI is a smartphone-enhanced structured pedagogy program where teachers are given upfront training on the science of reading, and equipped with high-quality, semi-scripted lesson plans linked with student workbooks, to run daily literacy lessons and given coaching. The program incorporates a digital layer; teachers and coaches assess children's literacy skills and are provided with integrated student tracking, coaching management tools, and training videos through a mobile app called SmartCoach. We study the effects of the program's literacy model, Inspiring Reading, on first-grade students in low-fee private schools in the Central Region of Ghana. We randomly assigned 80 schools to either the Inspiring Reading program or a control group during the 2024-25 school year, and measured outcomes using end-of-year Early Grade Reading Assessments (EGRAs). We complement the data from the assessments with the survey data.

The program causes large increases in student learning: our pre-specified primary outcome, overall EGRA reading scores, increases by 0.504 SDs ($p = 0.014$) which is the equivalent of more than two years of progress under the *status quo*. This puts TFLI at the 91st percentile of all reading interventions within the first of what will be three years of the program. The effects on individual components of the reading score are consistent with the program's theory of change. The largest effects are on mapping letters to sounds ($d = 0.757$, $p < 0.001$) and phonemic awareness ($d = 0.709$, $p < 0.001$), which are the key skills targeted by the program in grade 1. The impacts on oral reading fluency and reading comprehension are smaller and do not reach conventional levels of statistical significance, although they are quite large relative to typical impacts in the literature ($d > 0.2$). These are downstream skills that the program is building toward, and where effects are more likely as students progress through the program in grades 2-4.

Because the impacts are so large, we present a wide range of evidence that they are real, and not driven by statistical noise or issues with the exams we used. We pre-specified a single primary outcome and the exact data analysis that we would run, so the p-value for that main test can be interpreted literally. Our findings are also robust to a wide range of robustness checks that vary the controls and address the small level of school-level non-compliance with treatment assignments, with point estimates ranging from 0.44 to 0.52 SDs.[1] The attrition rate was just 20%, identical across study arms, and not differentially correlated with covariates by treatment status; we nevertheless compute Lee bounds and find a treatment effect range of 0.39 to 0.56 SDs, with both ends being statistically significant. Our exam scores come from an internationally standardized test (the Early Grade Reading Assessment); the NGO was blinded to the content of the test until the exams began, and the tests were administered by external contractors. We ran a separate experiment to estimate potential demand effects, randomizing whether each student was tested by an enumerator who was from the teaching profession. Assessor type matters for average scores but has no differential effects by study arm.

To understand how TFLI achieved such large effects so quickly, we develop a model of skill formation in which basic skills constrain the development of advanced skills. The model predicts that it is optimal to differentiate instruction: we should teach the basics

[1] Two schools from each study arm received the opposite of their assigned treatment status due to administrative errors. We find comparable impacts if we use the treatment they received instead of the one they were assigned (which was our pre-specified approach).

to children with lower skills and advanced materials to children who are higher-skilled. It also predicts that high teaching quality and targeted instruction are complements, so that programs that combine the two approaches (structured pedagogy and targeted instruction) will have particularly large impact. TFLI operationalizes this approach, providing scripted lesson plans and workbooks (to improve teaching quality) along with regularly-scheduled targeted instruction (for differentiation of instruction). Moreover, the SmartCoach app helps to enhance both components, easing the management of program adherence for coaches and also the process of assessment and differentiation for teachers.

Consistent with the model, we see the largest impacts on the most basic skills: letter sound knowledge and initial sound identification. More advanced skills progress by less. We also see stark variation in the effects of the program across the distribution of test scores. In particular, there are large and statistically significant reductions in the fraction of students who cannot recognize words or read any words in a passage, both of which fall by over 40%. Impacts at the higher end are smaller, which is consistent with the program building basic skills more at this grade level. The effects also appear to be larger for male students; control-group girls are ahead by 0.26 SDs, and the treatment closes ⅔ of this gender gap. This suggests that it may be more beneficial for weaker students more broadly. This would be consistent with the program's design, which focuses on supporting teachers in using assessment-informed instruction and in-classroom remediation.[2] Because we do not have baseline test score data, we cannot decompose our treatment effects by students initial test scores, but our other analyses are consistent with the prediction that weaker students and weaker skills are targeted more by the intervention.

We see evidence for a number of potential mechanisms for the treatment effects. A pre-specified index of teaching quality improves by nearly 2 SDs, with notable gains in key phonics activities such as have learners say the same correct sounds as the teacher and blending sounds to make words. Classroom observations also reveal increases in the teacher moving around the room and in student engagement with the workbooks. Student self-

[2] We do not see other evidence of gender-specific effects. The benefits are larger in schools with female principals; we see no evidence of differences in impacts by teacher gender or based on teacher-student gender match. This is consistent with the student gender pattern being driven by differences in skills rather than by other facts correlated with gender.

perceptions appear to improve, with statistically significant reductions in students thinking they are at the bottom of the class. Students are also more likely to practice reading at home, in line with previous evidence on shifting beliefs about relative performance (Dizon Ross 2020).

To further test these mechanisms we run an A/B test to examine how further enhancements to teaching quality affect test scores. A/B tests are rapid randomized experiments that allow organizations to improve their operations (Angrist, Cullen, and Magat 2025). We tested an intervention in which school leaders (principals) were trained to provide additional coaching to teachers on their implementation of the program, with the goal of improving teaching quality. Using a lower bar for statistical significance (which is standard in A/B testing) we see evidence of gains in quality from this intervention. The effects on learning are not yet distinguishable from zero, but based on these findings Inspiring Teachers is continuing the intervention in the 2025-26 school year. Moreover, the impacts on learning are quantitatively consistent with our model: we see larger relative effects on more-advanced skills, with the impact on reading speed being 70% of the main treatment effect, while the impact on letter sound recognition is just a 10% of the main effect.

Our model makes specific predictions about the program impacts we expect to see in second grade, which we will test in future data collection for the project. Specifically, we expect to see larger gains in more advanced skills now that students have developed the basic reading skills that constrain them. We also expect to see higher gains in advanced skills for students who are further up the skill distribution at the end of grade one, and lower gains for those who are at lower levels. We are currently tracking all the students from the initial sample that we selected at the beginning of grade one, and are planning to collect a second round of data on all students in June 2026. This will allow us to test these forecasts empirically.

The impressive gains achieved by TFLI have important policy implications because the intervention is scalable both over time and across space. We conducted a previous small-scale pilot RCT during the 2023-24 school year; the intervention ran in 4 randomly-assigned treatment schools that year for just four months. Comparing those treatment schools to 4 randomized control schools, we see gains of 0.25 SDs, with the impacts distinguishable from zero despite the small sample size. Moreover, the actual RCT results in 2024-25 are very close to what we would have extrapolated from this pilot based on the additional time spent in the program: the program ran for 2.25 times as long and had effects that were nearly 2.25 times larger. This suggests that continued exposure to the program may raise test scores almost linearly, so we can expect gains of over 1 SD by the time the program finishes at the end of grade three.

The intervention is also designed to be scaled across space, not just within Ghana but also across Africa. Within Ghana, Inspiring Teachers is already scaling up the program to more of the country and to different kinds of school. It is running in 139 schools in the 2026-27 academic year, including in 80 government schools. The organization has been invited to expand to 400 government schools and 100 low-fee private schools in 2026-27, and is collaborating with the national and regional offices of the Ghana Education Service to roll out TFLI in all 1,638 government schools in the Central Region by 2029-30. This expansion is slated to be highly cost-effective: the current marginal cost of the program is \$48 per student, and so the cost per 1-SD gain is \$96, which already makes the program competitive with existing interventions. By 2029, Inspiring Teachers' budget model predicts the cost will drop to \$6 per student, which would make it extremely cost-effective if its current effectiveness can be sustained.

Scaling TFLI across Africa more broadly will require adapting the materials to other local contexts, education systems, and languages of instruction. It has two key advantages on that front. First, today TFLI is English-language-first, which means that it can in principle be used across all of Anglophone Africa. It works even though English is not the native language of our study sample: TFLI has achieved significant gains in learning despite just 15% of our sample speaking English at home. This means it can serve as a complement to existing mother-tongue-first instruction programs. Second, TFLI's lesson plan developers use a component-based design system (where lessons are assembled from a common pool of adaptable components) and generative AI tools to accelerate lesson guide and workbook development. This allows Inspiring Teachers to efficiently leverage a highly scarce talent pool—highly-skilled instructional designers, which are rare not just in Africa but around the world. The organization is already using this tool to adapt the program to Uganda. They ran a preliminary pilot test of the program in Kanungu District during the 2025 school year, covering grade 1 classrooms. The pilot was not randomized, but they did post-intervention tests in both the program schools and in similar nearby schools. A regression-adjusted comparison of the mean test scores, following our specification for the main RCT, yields a difference of 0.514 SDs. These results suggest that the genAI-assisted curriculum adaptation approach can help the program scale to other countries with different early-grade reading curricula. Inspiring Teachers is in talks to do this in Zambia. TFLI has the potential to substantially narrow the learning gap between schools in Africa and those in the developed world.

Our results make contributions to three literatures in economics. First, we provide additional evidence that it is possible to drastically improve test-scores in learning-impoverished contexts. Previous work has shown that two types of intervention are capable of achieving impacts larger than half a standard deviation. The first is targeted instruction, which has proven benefits in a number of contexts (Duflo, Dupas, and Kremer 2011, Banerjee et al. 2007, Muralidharan, Singh, and Ganimian 2019) including in Ghana (Beg, Fitzpatrick, and Lucas 2023) and has been successfully scaled up (Banerjee et al. 2017). Angrist and Meager (2023) argue that targeted instruction has impacts of 0.9 SDs when implemented with high fidelity; fidelity (and, concomitantly, impacts) vary substantially across studies. The second is structured pedagogy, which has achieved large impacts in both local-level randomized trials (Piper et al. 2018*c*, Eble et al. 2021, Buhl-Wiggers et al. 2024) and at national scale (Piper et al. 2018*a*). It is also a key component of the extremely high-impact programs studied in Gray-Lobe et al. (2022) and Fazzio et al. 2021. We contribute to these existing findings by showing that large gains are achievable after just one grade of exposure, and using English-first instruction, despite most students speaking a different language at home.

Second, we show that targeted instruction and structured pedagogy can be combined successfully. While these two types of intervention are proven to have large impacts on their own, they have rarely been combined. Existing work on combining the two approaches uses observational data to show large impacts that are plausibly causal (Ibrahim et al. 2024). We build on this earlier work by randomizing the roll-out of an intervention that combines structured pedagogy and differentiated instruction, and also showing that this combination works in a totally different context. These findings are also part of a literature that studies complementarities between educational interventions (Mbiti et al. 2019, Kerwin and Thornton 2021, List, Livingston, and Neckermann 2011). We do not explicitly randomize the two aspects of the program, but our results, complemented by our theoretical framework, suggest that the two aspects of the intervention are complementary rather than substitutes.

Third, we also contribute to the theory of how differentiated instruction works. This is part of a broader literature about dynamic complementarities (Cunha and Heckman 2007). That literature has the feature that skills beget skills, sometimes called the "Matthew Effect". We build on this idea to develop a model in which basic skills constrain the development of more-advanced skills, which is a key assumption underlying the literature on "teaching at the right level", or TaRL (see e.g. Banerjee et al. (2017)). We build on this literature in two ways. First, we link it to work on differentiated instruction & TaRL, showing when it is optimal to teach to the bottom of the distribution versus the top. Previous theoretical work on differentiated instruction has taken as assumed that targeting instruction to a student's learning level is beneficial (Duflo, Dupas, and Kremer 2011); we derive that result from an underlying model of skill formation. Second, we contribute to the existing body of research for dynamic complementarities from randomized experiments (Bettinger et al. 2020, Shaikh 2025, Carneiro et al. 2025).[3] We complement that existing work by showing that our framework makes specific, testable predictions about the pattern of treatment effects that match what we observe in the data so far, as well as for what we should see in future rounds of data collection.

The remainder of the paper proceeds as follows. In Section 1 we describe the setting of the study and the TFLI intervention. We describe the data that we rely on in Section 2 and the empirical strategy we use to analyze it in Section 3. Section 4 presents the results, and Section 5 interprets them through a model of skill formation. In Section 6 we discuss evidence of the scalability of the intervention both over time and across space. Section 7 concludes.

[3] Shaikh reviews a number of studies that estimate models of dynamic complementarity using observational data, including Todd and Wolpin (2007), Aizer and Cunha (2012), Gilraine (2017).

1 Context and Intervention

1.1 Primary education in Ghana

In Ghana, primary education (known as basic education in the country) starts with two years of compulsory kindergarten, beginning at age four. These are followed by Basic 1 (B1), which typically starts at 6 years of age. B6 is the last year of primary school, and is followed by three years of three years of junior high school (JHS1-3). Students take the Basic Education Certificate Examination (BECE) at the end of JHS3; passing the BECE is required to enter senior high school (SHS), which runs from SHS1 to SHS3.[4] Primary schooling is delivered through both private and public schools. Public schools are free, while private schools include low-fee schools that are accessible to the poor as well as high-end schools that are much more expensive. According to Ministry of Education (Ghana) and UNICEF and Ministry of Education (Ghana) (2023), teachers in public schools are typically highly trained compared to teachers in private schools, especially at the basic level.

Ghana, like many low and middle income countries, has a high prevalence of low-fee private schools that offer primary education (Brion 2020). These schools are often concentrated in urban areas; some also exist in rural areas where the reach of public schools is limited. Low-fee private schools became more common in the wake of the inception of the Millennium Development Goals (MDGs), and aimed to fill the gap between the supply of public schools and the demand created by MDG 2, which called for free primary education for all. They also served to provide a choice to parents who were dissatisfied with government schools (Day Ashley et al. 2014). Typically, low-fee private schools in Ghana are independently-owned and run on a for-profit basis.

Although public school primary school teachers have better formal training than those at low-fee private schools, urban households often opt for the latter. This is mainly driven by perceived quality, closer supervision of teachers, better learning environments, and historically better learning and schooling outcomes (Day Ashley et al. 2014; Akaguri 2014). Another attraction of low-fee private schools is that they often use English, which is Ghana's official language, and the lingua franca of the country (Brion 2020). This is potentially a

[4] There is no required examination to proceed from B6 into JHS1.

selling point for parents because of its perceived economic returns and social status.

Government policy has attempted to promote the use of mother-tongue instruction (teaching students in the language they grew up speaking) from kindergarten through B3. However, the lack of clarity on language policy in the country as well as issues around teacher deployment, the large number of languages, parental demand for English, and the shortages of materials have rendered the implementation of this policy weak and uneven. These factors, together with others, have contributed to poor learning outcomes in early grade levels (Curto and Keane 2025). In practice, schools employ a mix of local languages and English in early grade instruction depending on the teacher's capacity and the availability of teaching and learning materials.

Enrollment has increased substantially in Ghanaian primary schools over the past few decades, primarily because of the Free Compulsory Basic Education (FCUBE) policy introduced in 1995 and the Complementary Basic Education (CBE) program implemented from 2012 to 2018 among others. In contrast, learning outcomes, especially at the basic level, have not improved by much. An Early Grade Reading Assessment implemented by the USAID across 168 districts in Ghana showed that at the end of second grade, pupils could read an average of 2.5 words per minute, with about 77 per cent of the students unable to read a single word (Social Impact, Inc. 2018, UNESCO 2023). This pattern of increasing enrollment but low progress on learning is common across much of the developing world (World Bank Group 2018).

A series of ambitious reforms have been implemented since 2017 to improve primary-school learning outcomes in Ghana. These include the development of teacher standards, a new curriculum, and new assessments, all with the aim of improving accountability and learning outcomes at the basic level of schooling. The standards-based curriculum introduced in 2019 emphasized foundational knowledge, including literacy and numeracy. Alongside this new curriculum are standardized tests at B2, B4, B6, JHS2, and SHS1 that are used to progressively test core competencies in literacy and numeracy (Ministry of Education Ghana). These tests are not used to determine student advancement, only to measure outcomes.

1.2 The TFLI Intervention

Tools for Foundational Learning Improvement (TFLI) is a smartphone-enhanced, structured pedagogy program designed to support teachers to deliver consistently high-quality early-grade instruction. We focus on the Inspiring Reading Program, which is the initial version of TFLI aimed at literacy skills. This program was approved by Ghana's National Council for Curriculum and Assessment and is aligned with the Ghanaian national curriculum and the Ghana National Teaching Council's continuing professional development points framework.

Each term, teachers receive a teacher guide containing daily lesson plans and a set of aligned student workbooks. Lessons follow a consistent pedagogical routine and are semi-scripted to enable them to be used in real time during classes. The student workbooks are designed to make learning visible, allowing teachers to monitor pupil responses during lessons, make in-the-moment instructional adjustments, and identify learners requiring additional support. See Appendix H for examples of pages from the teacher guide and student workbooks. The program features a digital layer, which teachers and coaches interact with through a smartphone app called SmartCoach.

The teacher guide, workbooks, and SmartCoach app support four integrated components: (i) upfront training in evidence-based literacy instruction; (ii) daily structured lessons supported by teacher guides and student workbooks; (iii) smartphone-based reading assessments and student progress tracking; and (iv) data-driven coaching and program management.

(i) Upfront Training in Evidence-Based Teaching

Teachers participating in the study received two days of upfront training prior to program implementation, followed by a one-day refresher training before each subsequent term (four days of training in total). Training introduced teachers to the twelve core pedagogical routines that underpin the program, as well as the "science of reading" principles that inform their use.[5]

The training model is designed to help teachers understand how each routine targets

[5] See Alvarez Marinelli et al. (2025) for an overview of the science of reading.

specific foundational reading skills. Teachers practice these routines in small groups to prepare for classroom implementation. Training is reinforced through short instructional videos embedded within the SmartCoach app, which teachers can access during the school term.

(ii) Daily Lessons Aligned with the Science of Reading

Instruction follows a consistent five-day instructional cycle. On Days 1–4, teachers deliver a one-hour structured literacy lesson grounded in the science of reading. Instruction integrates systematic phonics within a broader instructional sequence that progresses from oral language development to phonics, and subsequently to reading and writing.

Day 5 of each cycle is dedicated to assessment and remediation. Teachers administer a brief whole-class assessment aligned to the week's instructional content and provide targeted reteaching or additional practice based on pupil performance. These 5-day cycles are designed to match a school week, but also can be used on any day of the week in the case of school holidays.

(iii) App-Based Reading Assessments and Student Tracking

Each term, teachers conduct one-on-one oral reading fluency assessments with every pupil in their class using SmartCoach. During the assessment, pupils read a short passage aloud for one minute while the teacher records errors. The application automatically times the assessment and calculates reading speed and accuracy. SmartCoach aggregates these data to generate a class-level summary that is organized by reading proficiency, enabling teachers to monitor pupil progress over time and identify learners who are falling behind and may require targeted support.

(iv) Data-Driven Coaching and Program Management

Once the up-front training has occurred, school leaders and field staff use SmartCoach to conduct structured lesson observations and provide teachers with instructional coaching. The app includes observation checklists and decision-support tools that guide observers and generate targeted feedback for teachers, with the aim of making coaching more specific and actionable.

At the program management level, data captured through SmartCoach (including assessment completion, coaching activity, and lesson delivery) enables monitoring of implementation fidelity. This data is used to identify classrooms where program components are not being implemented as intended and to plan targeted follow-up support. During the 2024-25 school year that is the focus of this paper, this process was managed through a combination of spreadsheets and a database; subsequent iterations of the program have consolidated these functions within a web-based management dashboard.

2 Experiment and Data

Our data comes from a randomized trial in Ghana's Central region, centered around Cape Coast. Treatment ran in B1 (first grade) classrooms for the 2024-25 academic year. We collected data at the end of the intervention in June 2025 over a two week period. We did not collect any data at baseline beyond basic information about the schools.

Our sample was 80 low-fee private schools spread throughout the central region (Figure 1) selected by Inspiring Teachers based on interest in the program. The schools are independently owned; Inspiring Teachers does not own any schools. To be eligible, schools had to charge fees of 400 Ghanaian Cedis per term (1200 Cedis/year), which is about 5 percent of median household income (Ghana Statistical Service 2019). Of these, we randomly assigned 40 schools to receive the treatment in September of 2024. Because of challenges recruiting schools to participate in the intervention, and the necessity of beginning the program as close to the beginning of the school year as possible, we randomized batches of approximately 20 schools at a time over the course of a few weeks. Each batch was the 20 schools Inspiring Teachers was most easily able to contact and convince to join the program since the last batch. The randomization was stratified by batch and, within batch, by school size. We targeted a stratification cell size of 4 schools, following the best practice recommended by McKenzie (2025). Due to ties in the school size variable some cells had either 3 or 5 schools.

The randomization produced study arms that are balanced on baseline covariates[6]; randomization inference F-tests of overall balance following Kerwin, Rostom, and Sterck (2025) yield p-values of 0.60 for student-level characteristics and 0.55 for school-level characteristics (Table 1).

Four of the schools in the sample received the opposite of their assigned treatment status due to administrative errors. These arose due to a combination of the batched design of the randomization and the fact that many schools have extremely similar names. As a result, two of the control schools actually received the intervention and two of the treatment schools did not receive it. Our estimates use an intention-to-treat approach, analyzing the effect of the randomly-assigned treatment rather than the actually-received treatment. We also show that this does not make a substantive difference for our results.

In addition to the main experiment, we also conducted an A/B test among the 40 treatment schools, changing part of the program implementation to try to improve it.[7] This A/B test gave additional training to school leaders (principals) to enable them to provide coaching to the teachers in their schools, supplementing the coaching provided by Inspiring Teachers staff. 20 of the schools were assigned to receive the school leader coaching, and the other 20 were not.

Over the year the intervention ran, four schools closed down—two treatment and two control—so we had 76 total schools for our endline data collection. The 80 schools in our initial sample had 1,643 first-graders on their rosters at the beginning of the year. We were successfully able to find 1,322 students at endline, which is a 20% attrition rate. Attrition rates were not differential between treatment and control schools, and there is no evidence that patterns of attrition by baseline covariates differed by study arm (Table A1). We test for balance in this post-attrition sample using an expanded set of exogenous variables that we collected at endline. The post-attrition sample is balanced on the characteristics students (overall balance p-value = 0.20), teachers ($p = 0.92$), school leaders ($p = 0.52$), and schools ($p = 0.26$) (Appendix Tables A2, A3, A4, and A5).

[6] The difference is regression adjusted controlling for stratification cell fixed effects.

[7] Because this A/B test was altered how the program was run, we randomized the 40 schools that *actually* received the treatment to one of the two conditions, i.e., including two of the schools that were assigned to control status in the main study.

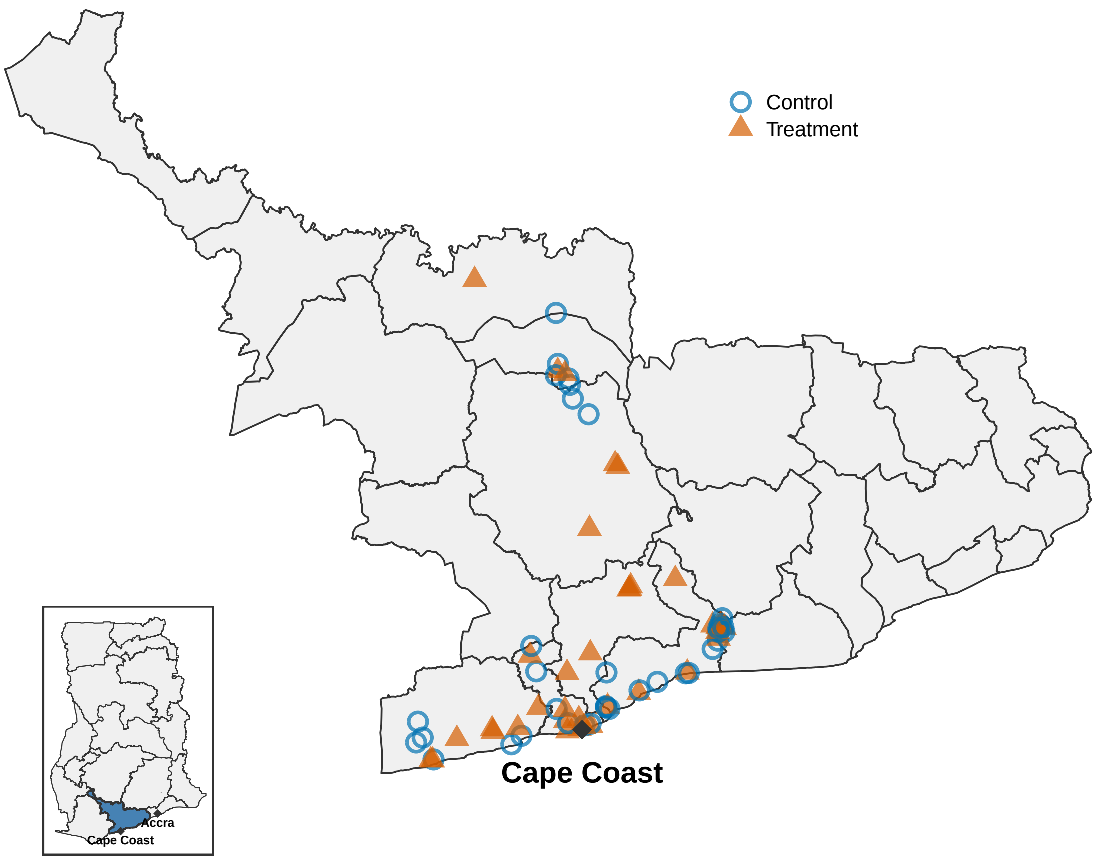
Table 1
Balance Table

	(1) Control Mean (SD)	(2) Treatment Mean (SD)	(3) Reg. Adj. Diff (T-C) (p-value)	(4) Obs.
Panel A: Student-Level Variables				
Male	0.541	0.537	0.003	1,416
	(0.018)	(0.019)	(0.897)	
Student Age (Years)	6.605	6.486	-0.117	1,381
	(0.035)	(0.036)	(0.229)	
Joint F-stat (omnibus, unadjusted)			0.66	
RI p-value (permutation)			0.60	
Panel B: School-Level Variables				
Total Number of Students	183.079	196.447	20.345	76
	(20.743)	(29.964)	(0.201)	
Number of Teachers	11.000	11.794	0.706	71
	(1.019)	(1.346)	(0.556)	
Number of BS1 Students	22.757	19.514	-4.400	74
	(2.117)	(2.342)	(0.101)	
Proportion Male	0.544	0.548	0.017	74
	(0.022)	(0.026)	(0.638)	
School Fee (GHS)	174.207	201.000	27.547	59
	(18.720)	(18.167)	(0.317)	
Joint F-stat			0.82	
RI p-value			0.55	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year. Joint F-statistic based on (Kerwin, Rostom, and Sterck 2025). Differences in column 3 are estimated using a linear regression that controls for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

In each first-grade class, enumerators ran early grade reading assessments (EGRAs), surveyed the students, teachers, and school leaders, and filmed one literacy lesson. We hired the 23 enumerators specifically for the endline data collection; they had no previous connection to Inspiring Teachers. Three of the enumerators were school improvement support officers (SISOs)—Ghana Ministry of Education staff hired with the intent of working with Inspiring Teachers in future years to support implementation of the intervention in government schools. 12 of the enumerators had experience as teachers or had worked in the education sector, including all three of the SISOs. To mitigate potential concerns about

Figure 1
Study Sample Schools



different enumerator backgrounds (external evaluator vs. stakeholder, education experience vs. no experience) impacting EGRA evaluations, we randomized enumerators to schools and students to enumerators within a school, finding no evidence that this matters for our results.

To further ensure validity of the EGRA tool and prevent teaching-to-the-test, we imposed strict controls the contents of the examination. The only member of the Inspiring Teachers staff that saw the EGRA contents before endline data collection were the SmartCoach development team who had to program the assessment into the app. We required these development team members sign a non-disclosure agreement to maintain confidentiality of the questions. In particular, co-PI Simon Graffy, who directs the NGO, did not have access to the test questions.

2.1 Measure of Reading Ability – EGRA

We used an English language EGRA version to match the language of instruction. We included eight standard subtests to capture different reading skills, including listening comprehension, letter names, letter sounds, initial sound identification, familiar word reading, non-word reading, oral reading fluency, and reading comprehension. Our pre-specified, primary outcome is a combined index of all eight subtests. We constructed the index by taking the first principal component of the control-group data and applying those weights to the treatment group as well.[8] We specified the exact data-driven procedure for constructing the weights in the analysis plan we posted in advance of the beginning of data collection. We also report results for each of the subtests individually.

We arranged the subtests in order of increasing complexity. Listening comprehension does not require any ability to read, so it was first. In this subtest, the enumerator read a short passage to the student then asked them three questions about what happened in the story. Letter names and letter sounds are the next step in the complexity ladder: students don't have to be able to understand full words, but need to be able to recognize various properties of letters. For these subtests, enumerators showed students a grid of 100 letters then asked the student to tell them the name of the letter or what sound it makes respectively. Slightly more complex is "phonemic awareness" which asked students to identify the first letter sound in a series of 10 words.[9] This is more challenging than looking at letters alone because students can be confused by the extra sounds in the word.

The next rung is for students to be able to combine letters and read words, so we tested familiar-word and non-word reading. Similarly to the letter subtests, students had a minute to read off words from an 50-word grid. Familiar-word reading included a mix of words

[8] Table A8 shows the weights that the index puts on each subtest. Consistent with our understanding of the development of early reading skills, listening comprehension and letter names get the lowest weights, while the highest weight is is on oral reading fluency.

[9] E.G. for the word "up" the correct answer is /uh/.

students could sound out (or decode) using known phonics sounds[10] and also sight words that do not follow standard phonics rules.[11] Non-word reading tests only phonics skills. Non-words are collections of letters that do not form actual words, but can be sounded out using basic phonics rules.[12] Nearing the top of the complexity ladder was oral reading fluency. In this subtest, students were given a simple, short story of 56 words to read. This is more complex than familiar- and non-word reading because words are now strung together in a cogent order. This means errors can be serially correlated–if a student can't read a word early on in a sentence, it may make it more challenging to read the rest of that sentence. Finally, the top of the complexity ladder is understanding what you read. In this final subtest, after reading the oral reading passage, students were asked five simple questions about what happened in the story, which the enumerator marked correct or not. These eight subtests give us a complete picture of a student's reading ability.

The specific EGRA we used was an adaptation of a previous exam run by the Ghana Education Service and RTI International in 2015. We supplemented this with modules used by Kerwin and Thornton (2021) in a large literacy intervention in Uganda in 2013. We used existing, vetted EGRA materials for several reasons. First, this boosted our efforts to avoid teaching-to-the-test. Since the materials were fully external and developed prior to our intervention there was no chance the materials could be contaminated by the program. Second, because these materials had been used successfully in other research, this helped us avoid floor or ceiling effects that could limit our ability to detect the effects of the program.

The EGRAs were administered in-person, in a one-on-one setting by outside enumerators between June 16 to 30 2025. The enumerators scored the EGRAs using SmartCoach, but used laminated sheets with the letter/word grids and story on them for the students to read from. The schools were aware that they were part of a study of the Inspiring Reading program, but the enumerator teams did not identify themselves as working for Inspiring Teachers. If asked, they were trained to explain that they do not work for the organiza-

[10] E.G. "map" is /mmmm/ /ă/ /p'/.

[11] E.G. "said" is pronounced /s/ /e/ /d/ rather than the typical sound of "ai", /ā/. Many sight words actually follow more advanced phonics rules, so advanced readers could sound them out using phonics, but early readers could not.

[12] E.G. "gak" would be pronounced /g/ /a/ /k/. During enumerator training, we emphasized teaching the correct pronunciation of these non-words and how to recognize common errors students might make.

tion and that they were collecting data for a study being conducted by the University of Washington and the University of Ghana.

We employed an EGRA trainer with extensive experience training enumerators in Sub-Saharan African settings run the EGRA-focused part of our endline data collection training. We required enumerators to achieve a minimum standard of inter-rater reliability with known correct test answers before they were hired.

2.2 Measure of Teaching Quality

To gain an understanding of how the intervention directly affected teaching quality in grade one literacy classes, we had enumerators observe and film a literacy lesson during the school visits. We had the enumerators fill out the observation tool that Inspiring Teachers uses to track the progress of teachers in the program. The tool is similar to the World Bank's Teach Primary tool (Molina et al. 2018). Enumerators successfully observed a literacy lesson at 65 of our 76 schools; the 11 schools without observations were due to absent teachers, time constraints, and battery issues with the smartphones they used to do the recordings.

The tool measures both the quality of instruction, and adherence to the program. The key metrics for quality center around the learning climate, the nature of instruction (e.g., does the teacher use active learning techniques?), and whether the teacher regularly checks if students understanding the material. More specifically, the tool grades teachers on their actual instructional practices. For example, did the teacher actively model the phonics sound of the day? (and did they do so correctly?). The enumerators also observed student behavior, measuring if the students appear to be engaged, involved, and actually learning material during the class.

To measure adherence, the tool checks if teachers included all of the components during the lesson. These components are oral language, phonics, reading, and writing, which is the order and structure of a class period as prescribed by the intervention. Thus, the tool is opinionated toward what it considers "good" teaching. Beyond this, we also measured what proportion of the suggested number of lessons were actually run, what proportion of suggested assessments were run, whether the class used workbooks, whether the teacher used a teaching guide, and if the teacher sent home report cards. Some of these components can only be calculated for the treatment group because e.g. the control group does not have a suggested number of lessons. Through this we collect data not just on how well the treatment group complies with the intervention program, but also how different this is in the control curriculum.

2.3 Survey Measures

In the student survey, participants provided information on the demographic and socio-economic status of their households. Other sections of the survey cover learning behavior and study habits, academic self-perception, motivation, and academic and career aspirations. We also elicited information on perceptions of the classroom environment, the clarity of instruction, and the safety of the school environment.

The school leader survey measured demographics and professional experience for school leaders as well as their access to and use of digital tools, including smartphone ownership and use of the internet. We also asked about confidence levels in leading the coaching of teachers, and had the school leader report information about school enrollment and student performance and learning outcomes.

The teacher survey covered similar topics to the school leader survey. It also asked questions about to teachers' participation in and implementation of the TFLI program and their assessment and instructional practices. The survey was administered to first-grade teachers whose students took the EGRA tests.

3 Empirical Strategy

Our empirical strategies rely on the random assignment of schools to the treatment or control group. Our primary outcome, regression equation, and inference method were all fully pre-specified in our analysis plan.[13] We run the following regression to analyze the data from our main experiment:

$$Y_{ij} = \beta_0 + \beta_1 TFLI_j + Z_j'\tau + X_i'\gamma + \epsilon_{ij} \tag{1}$$

[13] https://www.socialscienceregistry.org/docs/analysisplan/9219

In this equation, Y_{ij} is the outcome of interest and i indexes students, which are nested within their original schools indexed by j. $TFLI_j$ is the indicator for a school being randomly assigned to receive the TFLI program. Z_{ij} is a vector of indicators for the stratification cells used in the lottery that assigned schools to study arms. X_i is a vector or control variables. We include the following pre-specified variables as controls: an indicator for being male, indicators for each value of age in years (at the beginning of the academic year), and the interactions between the two. We winzorize age at the 5^{th} and 95^{th} percentiles. We replace missing values of age or gender with separate categorical values, and include those when building the categorical indicators and interactions, so that these missings are dummied out in a fully nonparametric way.

de Chaisemartin and Ramirez-Cuellar (2024) note that in stratified experiments with small strata, the power and size of hypothesis tests are optimized by clustering inference at the level of a stratification level rather than a school. We follow their guidance in constructing our standard errors, since we have an average cell size of four schools. All p-values for our main analyses are based on randomization inference with 1,000 permutations. We show via a set of *ex ante* simulations that randomization inference gives the same p-values as clustering standard errors at the level of the stratification cell; these simulations informed our analysis plan.

Our primary outcome is an index of all the subtests from the EGRA assessment constructed using the first principal component of the control-group data and applying those weights to the treatment group as well. Because this is our only primary outcome, we do not correct for multiple hypothesis tests. We report results from each subtest individually, but in line with our analysis plan we do not correct for multiple testing for these secondary outcomes. For the EGRA index, and each subtest, we report results in each test's natural unit (SDs for the index, the number correct per minute for every subtest besides listening and reading comprehension, and the number of correct answers for the comprehension questions), and also in Equivalent Years of Schooling (EYS). Typically EYS are calculated by rescaling the treatment effect by the progress from baseline to endline in the control group, but we did not run a baseline survey. We use the conversion factor from Evans and Yuan (2019) for Ghana to calculate EYS from our estimates.

We also use Equation 1 to estimate effects from the student and teacher surveys including students' perceived class rank, and career and academic aspirations.

In Appendix A, we show additional estimates without baseline controls, without stratification cells, and using the received treatment status rather than assigned treatment status[14]Table A6. Although we do not have evidence of differential attrition (Table A1), we show Lee (2005) bounds in Table A7.

3.1 Enumerator Demand Effects

To test for and partial out any potential enumerator demand effects, we estimate the following regression. Because we randomized each enumerator-student assessment pairing, β_1 identifies the causal effect of being assessed by a particular type of enumerator.

$$Y_{ije} = \beta_0 + \beta_1 ET_e + Z_j'\tau + X_i'\gamma + \epsilon_{ije} \tag{2}$$

where everything is the same as Equation 1 except the following changes. e indexes the enumerator, and ET_e is an indicator for an enumerator being of a particular type. We separately consider two indicators for enumerator type: 1) if an enumerator has teaching experience or not, and 2) if they are a SISO or not. We consider these two types in two separate regressions, and not together, because 100% of SISOs have teaching experience.

We report multi-way clustered standard errors at the enumerator-school level. Each enumerator who visits a school defines a separate cluster; if two enumerators visit the same school, they form two distinct clusters. While each student is assessed by only one enumerator, the randomization design creates dependencies across all enumerators within a school that affect the test score variance structure. This clustering approach accounts for these dependencies. We validated this inference strategy via ex ante simulations and pre-specified it in the analysis plan.

We do not use randomization inference for Equation 2. This is because randomization inference tests the sharp null that the treatment effect of having an enumerator of a given type is zero for everyone, which is unlikely to hold in this context. It may be true that the

[14] Two schools had treatment status swapped during implementation.

average effect of being assigned an enumerator with teaching experience is 0, but since each enumerator is a different person, it is likely that within enumerator types, each enumerator has varying effects on test scores based on skill or experience running EGRAs, which would mean randomization inference would always reject the null even if the true average effect of being assigned a given type of enumerator is zero. Indeed, we can test this in our data and can easily reject the joint null hypothesis that all enumerator effects are zero for enumerators with teaching experience (p=0.01), and with no teaching experience (p=0.00). We cannot reject this null for SISOs (p=0.49) since only three enumerators are SISOs.

We also estimate treatment effect heterogeneity by enumerator type using the following regression.

$$Y_{ije} = \beta_0 + \beta_1 TFLI_j + \beta_2 ET_e + \beta_3 TFLI_j \cdot ET_e + Z_j'\tau + X_i'\gamma + \epsilon_{ije} \quad (3)$$

where the coefficient of interest is β_3, which identifies the differential effect of the TFLI treatment depending on the type of enumerator who ran the assessment.

We cluster standard errors at the stratification cell and at the treated-enumerator-school level. Each treated enumerator (either experienced teachers or SISOs, depending on the specification) who visits a school defines a separate cluster. The randomization design creates dependencies across treated enumerators within schools that affect the outcome variance structure. This clustering approach accounts for these dependencies. We validated this inference strategy via *ex ante* simulations and pre-specified it in our analysis plan.

3.2 Quality and Compliance Effects

To test for teaching quality and compliance with the TFLI program, we construct the pre-specified indices described in Section 2. We analyze these in several ways. First, we use them as the outcome variable Y_{ij} in Equation 1. We also estimate the following regression using two stage least squares using $TFLI_j$ as an instrument for $Compliance_j$.

$$Y_{ij} = \beta_0 + \beta_1 Compliance_j + Z_j'\tau + X_i'\gamma + \epsilon_{ij}$$

This lets us estimate how well the program would have worked with full quality/compliance, under the assumption that the effects of the randomized intervention operate only through compliance with the program.

The measures we use for teaching quality and compliance are generally measures of teacher behavior, but we collect test scores for each student. This gives us two different levels of aggregation at which to estimate treatment effects: student-level scores and teacher-level means. Aggregating to the teacher level discards much of the variation in test scores (and student-level covariates that can explain some of it), but leaving the data at student level implicitly weights treatment effects by class size. To reconcile this, we report results at both levels of aggregation for these variables.

4 Results

Our pre-specified primary outcome is the overall EGRA reading score. TFLI improves this measure by 0.504 SDs (p=0.014), equivalent to 2.2 years of status-quo instruction using the conversion factor from Evans and Yuan (2019)[15]. Panel A of Table 2 presents effects on all EGRA components.

The treatment effects align with our theoretical framework and the program's emphasis on foundational phonics. Letter sound knowledge and initial sound identification increase by 0.76 and 0.71 SDs (both $p < 0.001$). Non-word reading increases by 0.547 SDs ($p < 0.001$). These three core phonics skills drive the overall effect. More advanced skills show smaller gains. Familiar word reading and oral reading fluency increase by 0.31 and 0.20 SDs respectively, neither significant at conventional levels, although the former is quite close to the cutoff of 0.10. Reading comprehension increases by 0.25 SDs (p=0.27). This gradient from basic to advanced skills matches both our model's predictions and the program's first-grade focus on foundational literacy.

Listening comprehension and letter names are not skills emphasized by TFLI and show the smallest treatment effects (0.175 and 0.145 SDs). Letter name knowledge is also a skill

[15] Evans and Yuan report that students gain $\tilde{0}.22$ SDs each year in Ghanaian status quo literacy instruction, so to convert to this equivalent years of schooling, we divide the effect size in SDs by that amount.

emphasized in the status-quo curriculum. These skills, while important, stand farther from the phonics ladder of skills so it is unsurprising that students show smaller gains here.

These learning gains are very large even relative to those of other successful education programs, so we present a wide range of robustness tests to show they are real. Table A6 shows various different specifications where we vary using controls for age, sex, and stratification cell fixed effects. Because of an administrative error, the treatment statuses of four schools were reversed. In the even columns, we define treatment as actual receiving TFLI (as opposed to being randomly assigned to treatment) while the odd columns define treatment in the standard way. Finally, during the course of the academic year, four schools closed in one stratification cell, including both treatment schools. With no variation in treatment status, that cell is dropped from our regressions.[16] In columns 7 and 8, we pool the remaining school from that stratification cell into the cell with the next fewest amount of students, so the school's data is retained by OLS. None of these specifications change our conclusions. The overall treatment effect estimates vary between 0.44 and 0.52 SDs, and all are significant at at least the 0.05 level.

Similarly, although we find no differential attrition across treatment arms (Table A1), Table A7 presents Lee bound estimates for the overall reading score and all EGRA subtests. The upper and lower bounds for the overall index are both positive and significant at the 0.01 level, and are 0.560 and 0.394 respectively. This is also true for the subtests that showed the largest gains in our main analysis (letter sounds, initial sound identification, and non-word reading). For all other subtests besides familiar word reading (which did not have significant point estimates in our main specification), neither the upper nor lower bounds are significant. The upper bound for familiar word reading becomes significant at the 0.1 level.

SISOs do give systematically higher EGRA scores (Table D1), while teachers do not Table D2); neither pattern is systematically higher in the treatment group, and adjusting for enumerator type leaves our treatment effect estimates almost unchanged.

[16] `reghdfe` drops singleton cells explicitly (Correia 2016); OLS includes them but they do not contribute to the estimates. The two approaches produce numerically identical results with our data.

Table 2

Causal Effects of the Intervention on Reading Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
Panel A: Overall Scores																		
Treatment Effect	2.219	0.504	0.161	0.175	2.892	0.145	13.371	0.764	2.677	0.712	4.265	0.312	4.073	0.547	4.553	0.213	0.233	0.251
S.E.	(0.945)	(0.215)	(0.146)	(0.159)	(3.065)	(0.153)	(3.048)	(0.174)	(0.566)	(0.151)	(2.768)	(0.203)	(1.325)	(0.178)	(4.107)	(0.192)	(0.183)	(0.198)
R.I. p-value	[0.014]**		[0.249]		[0.256]		[0.000]***		[0.000]***		[0.125]		[0.001]***		[0.263]		[0.269]	
Control-group values																		
Mean	-0.000		0.540		32.660		17.459		3.870		8.331		2.974		13.169		0.381	
SD	4.400		0.920		19.980		17.496		3.760		13.668		7.441		21.426		0.926	
Adjusted R^2	0.175		0.143		0.105		0.220		0.219		0.094		0.114		0.080		0.064	
Panel B: Zero Score Students																		
Treatment Effect (% Change)	-0.007		-0.072		-0.005		-0.180		-0.200		-0.164		-0.316		-0.153		-0.114	
S.E.	(0.010)		(0.067)		(0.013)		(0.063)		(0.060)		(0.090)		(0.073)		(0.097)		(0.071)	
R.I. p-value	[0.485]		[0.232]		[0.624]		[0.003]***		[0.000]***		[0.034]**		[0.000]***		[0.038]**		[0.126]	
Control-group values																		
Mean	0.020		0.689		0.049		0.309		0.363		0.396		0.726		0.349		0.819	
SD	0.139		0.463		0.215		0.462		0.481		0.489		0.446		0.477		0.385	
N (# students)	1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298	
C (# stratification cells)	21		21		21		21		21		21		21		21		21	
Adjusted R^2	0.064		0.131		0.025		0.117		0.144		0.105		0.152		0.118		0.080	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Panel A shows the overall scores. Panel B shows the change in the proportion of zero scores which are defined as a binary variables that equals 1 if a student scores 0 on that EGRA component. Overall Reading PCA index is a weighted average of all the other components, where the weights correspond to the first principal component of control-group test scores. EYS stands for Equivalent Years of Schooling and is equal to the treatment effect in SDs divided by 0.22 (Evans and Yuan 2022). CLPM is correct letters per minute and CWPM is correct words per minute; both are calculated as the score on the respective subtest divided by the time taken. SDs are measured in control-group standard deviations. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). Randomization-inference p-values, clustered by school and stratified by stratification cell and using 1,000 permutations, in square brackets [ ]: *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

4.1 Distributional Effects

There is broad concern that successful education programs only benefit strong students while providing no benefit for weaker ones (Rudalevige 2003). In this section, we present evidence that TFLI does not exhibit this behavior.

Panel B of Table 2 shows the reduction in the number of students who scored zero on each subtests. see large improvements for these students at the bottom of the distribution across almost all the subtests including those that did not have significant average effects. The exception is the overall reading index, because there are nearly no students who score zero on that. A zero score on the overall reading score (i.e., the lowest value on the PCA index) entails scores zero on every subtest in the EGRA. This is only true five students out of the 1,298 members of our sample. The number of students who could read none of the words in the passage decreased by 44%. Other phonics skills show similarly large increases which all significant at the 0.01 or 0.05 levels. Reading comprehension does not have a statistically significant increase, but the magnitude of the change in non-readers is still economically significant with a 14% increase. Similar to our main results, listening comprehension and letter names are largely unaffected. Overall, there are large improvements at the bottom of the distribution across most skills.

Table 3 shows further evidence that TFLI benefits weaker students more broadly. In the control group, girls are about 0.26 SDs ahead of boys on aggregate reading skills. Panel A shows that treatment increases test scores by 0.205 SDs ($p = 0.054$) more than for male than female students, closing ²⁄₃ of the gender gap. Panels B and C show that this is not driven by other gender-specific effects of the program. Students with female teachers benefit more from treatment by ~ 2 SDs, but this is not significant at conventional levels. The difference is also likely driven by gendered selection into becoming a teacher; only 10.5% of the teachers in our sample are men. There are no significant differences by school leader gender or by gender match between students and their teachers or school leaders. This is consistent with TFLI benefiting male students more because they have weaker reading skills rather than because of other gender dynamics in the program.

A different way to consider distributional effects is through quantile regressions (Koenker and Xiao 2003), and distribution regressions (Chernozhukov, Fernández-Val, and Melly 2013). We briefly give an intuitive explanation of the difference between these two methods then apply them to this study.[17]

In Figure 2, we plot the survival functions (1-CDF) of letter sound knowledge separately for the treatment and control groups. These two CDFs visually summarize all differences in the treatment and control distributions (without controls). For example, because the treatment CDF is above the control CDF at every point, TFLI had a positive treatment effect at every level of letter sound knowledge.[18] Similarly, since the treatment is always to the right of the control CDF, TFLI has a positive treatment effect at every quantile. Quantile and distribution regressions allow us to statistically analyze these differences.

Quantile regressions estimate the difference in letter sound knowledge for students at a particular quantile. For example, they can estimate how much more letter sound knowledge the median treated student has than the median control student. In Figure 2, this is the green horizontal line connecting the treatment and control distributions. Quantile treatment effects (QTEs) are the horizontal difference between CDFs. Distribution regressions, on the other hand, fix a threshold of letter sound knowledge and estimate how much larger the proportion of treatment students who have at least that much letter sound knowledge is than the proportion of control students. In Figure 2, this is the purple vertical line. Distribution regressions are the vertical difference between survival functions.[19]

So far, this discussion has only considered unconditional estimates of QTEs and distributional effects, but both methods allow for controls. When control variables are included in quantile regression, the estimated treatment coefficient represents a weighted average of conditional quantile treatment effects across the distribution of covariates.[20] The weight assigned to each covariate value is proportional to both the marginal probability of that

[17] Our explanation expands on the explanation given in Kook and Pfister (2025).

[18] This is not necessarily equivalent to having a positive treatment effect for each individual student, which is unobservable; see Buhl-Wiggers et al. (2024) for a detailed discussion of this issue and potential ways to solve it.

[19] Distribution regressions can also be used to estimate the difference in treatment vs control proportions *below* a threshold, in which case the estimate is given by the vertical difference between the survival functions times negative one.

[20] Both methods assume constant treatment effects across the covariate distribution without inclusion of interaction effects.

covariate value and the conditional density of the outcome variable at the τ^{th} quantile given that covariate value. This weighting scheme implies that covariate values where observations cluster more densely around the quantile of interest receive greater weight in the estimated average effect. The same is true for distributional regressions, but the conditional density is at the threshold of interest rather than the quantile (Angrist, Chernozhukov, and Fernández-Val 2006).

Figure 3 shows QTEs and distribution regression effects for basic phonics skills, letter sound knowledge, and initial sound knowledge. In panels A and B, both of these basic skills show large treatment effects throughout their distributions. This is consistent with our model which predicts that basic skills should improve for all students because basic skills are unconstrained by any pre-requisite skills. Panels C and D show QTEs for the same skills. For letter sound knowledge, QTEs increase monotonically across quantiles. This is not inconsistent with TFLI benefiting students thought out the skill distribution. Combined with the distribution regression results in Panel A, this pattern indicates that while TFLI improved outcomes throughout the distribution, treatment helped a larger *proportion* of students cross low thresholds (distribution effects) even as higher-performing students experienced larger *absolute* score gains (QTEs).

There are similar patterns for more advanced skills. Figure 4 shows distribution regressions and QTEs for familiar- and non-word reading. Panels A and B show that most of the improvement happens at the bottom of the distribution with slightly more improvement higher up in the distribution for non-word reading. This is also true for the most advanced skills (oral reading fluency, and reading comprehension) in panels A and B of Figure 5. For all of these advanced skills, the QTEs show that higher performing students experience larger gains.[21] Overall, TFLI benefits a higher fraction of weaker students as compared to more-advanced ones, which is consistent with our model of skill formation.

[21] We omit the QTEs for reading comprehension because that subtest has only five points of support.

Table 3

Treatment Effect Heterogeneity by Gender

	(1) Equiv. Yrs. of Schooling	(2) SDs	(3) Equiv. Yrs. of Schooling	(4) SDs
Panel A: Student Gender				
Treatment Effect	1.735	0.394		
SE	(0.925)	(0.210)		
RI p-value	[0.070]*			
Treat × Male Student	0.902	0.205		
SE	(0.456)	(0.104)		
RI p-value	[0.054]*			
N (# students)	1,298			
C (# stratification cells)	21			
Proportion Male Students	0.550			
Panel B: Teacher Gender				
Treatment Effect	3.528	0.802	3.636	0.826
SE	(0.721)	(0.164)	(0.723)	(0.164)
RI p-value	[0.032]**		[0.020]**	
Treat × Male Teacher	-9.092	-2.066	-8.899	-2.023
SE	(2.486)	(0.565)	(2.487)	(0.565)
RI p-value	[0.214]		[0.209]	
Treat × Teacher-Student Match			-0.415	-0.094
SE			(0.397)	(0.090)
RI p-value			[0.398]	
N (# students)	1,227		1,204	
C (# stratification cells)	21		21	
Proportion Male Teachers	0.105		0.105	
Panel C: School Leader Gender				
Treatment Effect	3.137	0.713	3.131	0.712
SE	(1.933)	(0.439)	(1.983)	(0.451)
RI p-value	[0.415]		[0.383]	
Treat × Male School Leader	-0.230	-0.052	-0.431	-0.098
SE	(2.203)	(0.501)	(2.181)	(0.496)
RI p-value	[0.976]		[0.940]	
Treat × Student-Leader Match			0.365	0.083
SE			(0.530)	(0.121)
RI p-value			[0.498]	
N (# students)	1,247		1,220	
C (# stratification cells)	21		21	
Proportion Male School Leaders	0.776		0.776	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Figure 2

Quantile vs Distribution Regression for Letter Sound Knowledge

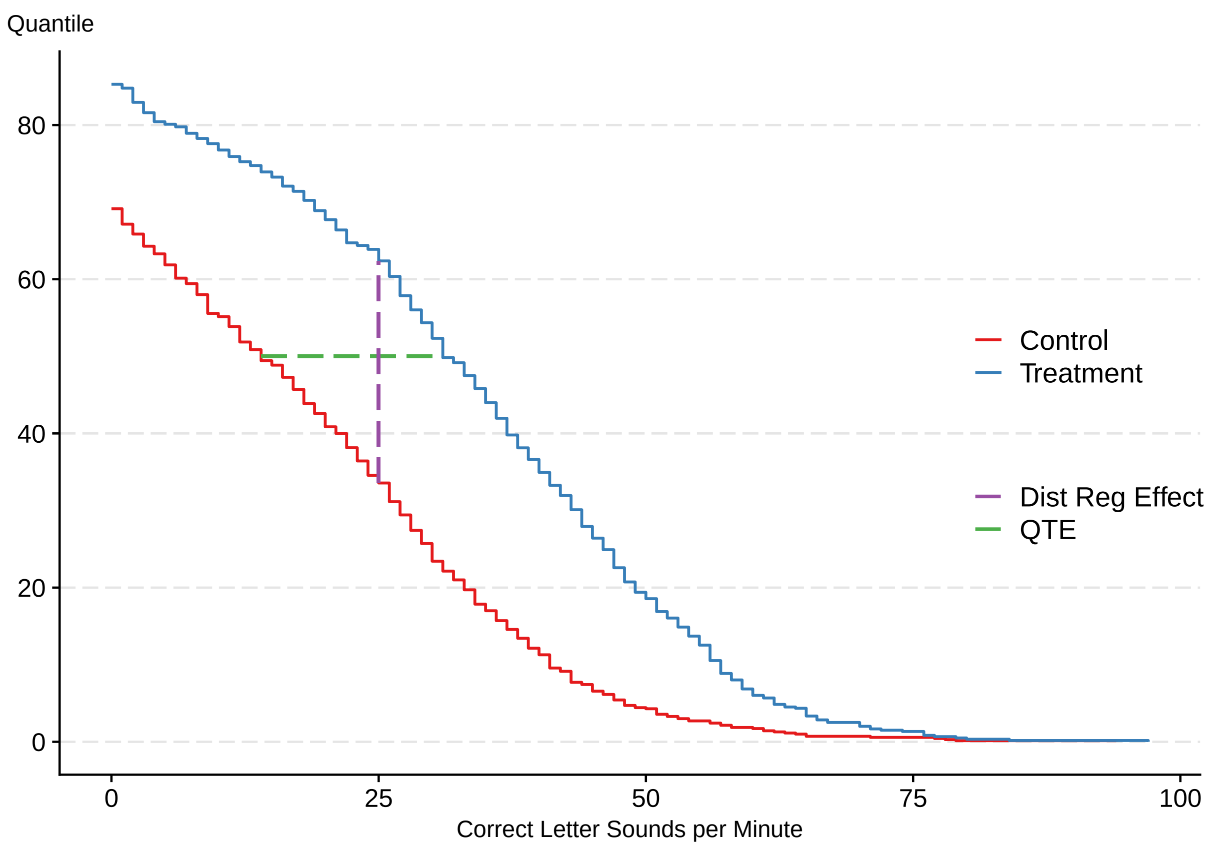
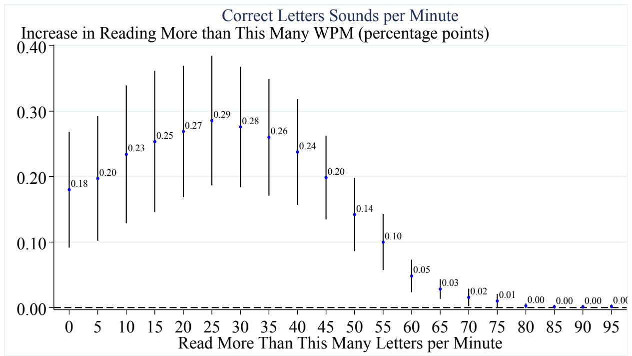
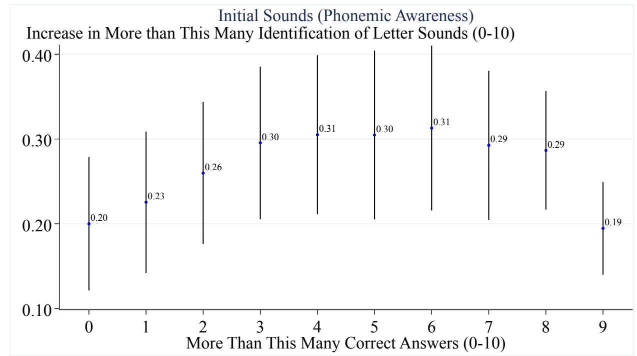


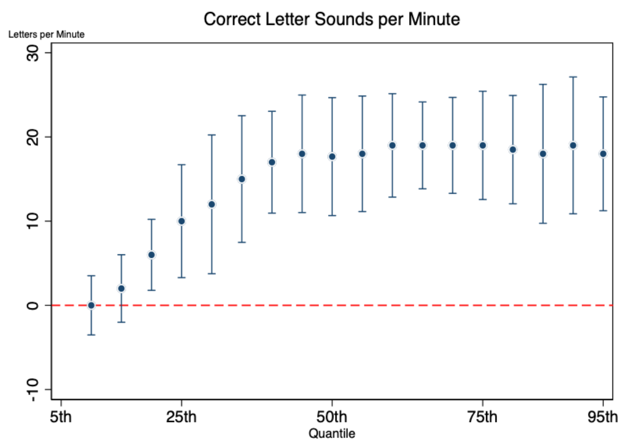
Figure 3
Basic Skills Distribution Effects



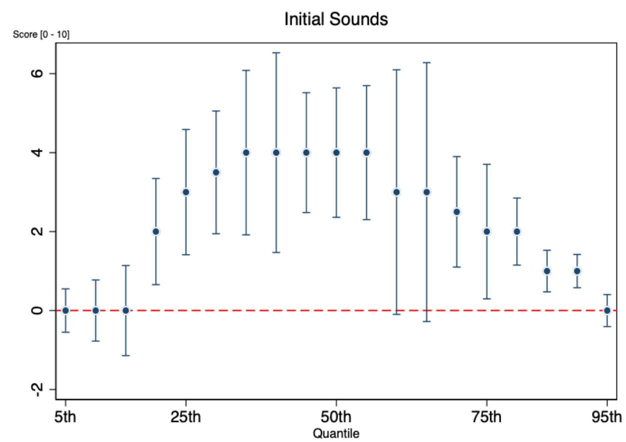
Panel A: Distribution Regression Effects for Letter Sound Knowledge



Panel B: Distribution Regression Effects for Initial Sound Knowledge



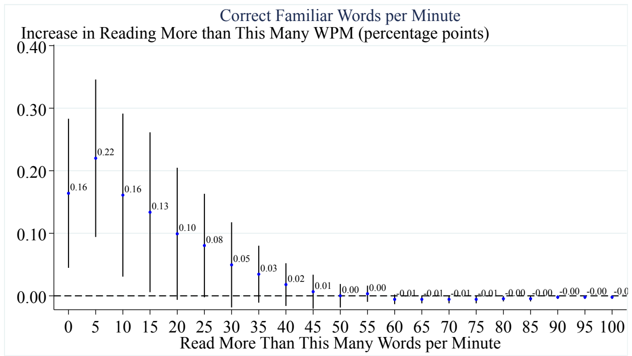
Panel C: QTEs Effects for Letter Sound Knowledge



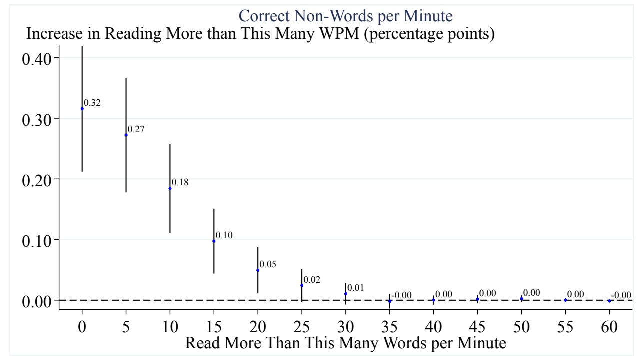
Panel D: QTEs Effects for Initial Sound Knowledge

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

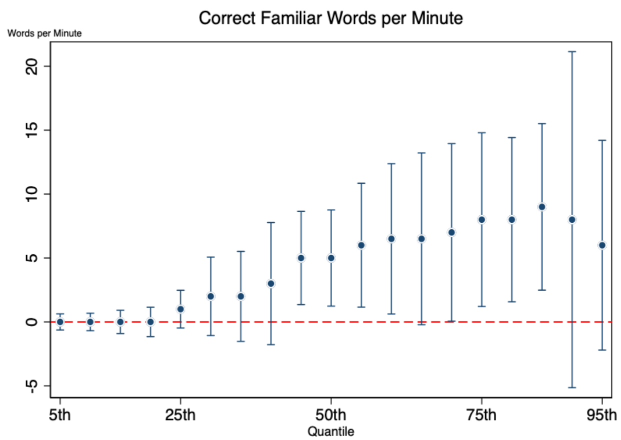
Figure 4
Word Reading Distribution



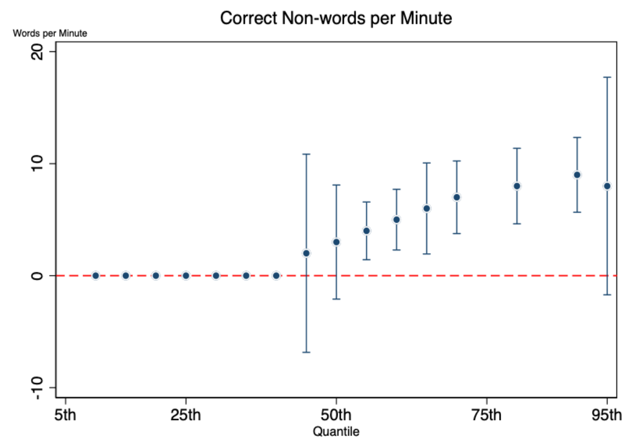
Panel A: Distribution Regression Effects for Familiar Word Reading



Panel B: Distribution Regression Effects for Non-Word Reading



Panel C: QTEs Effects for Familiar Word Reading

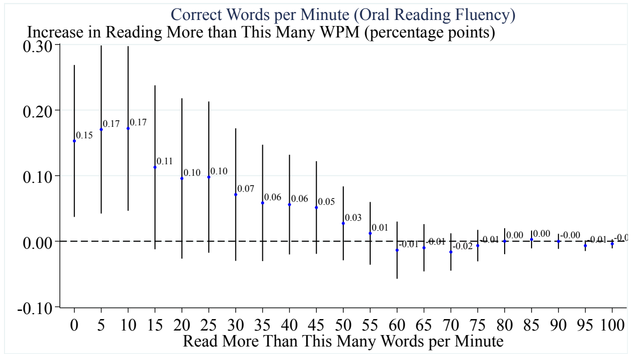


Panel D: QTEs Effects for Non-Word Reading

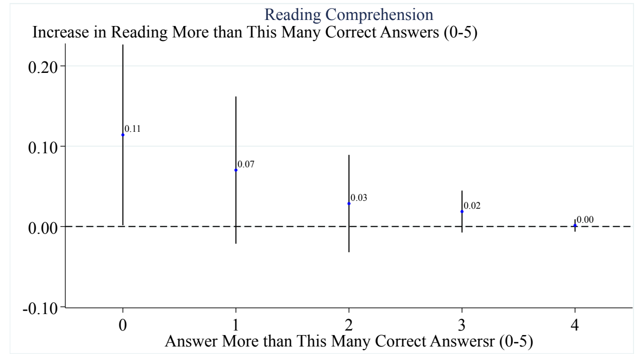
Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Figure 5

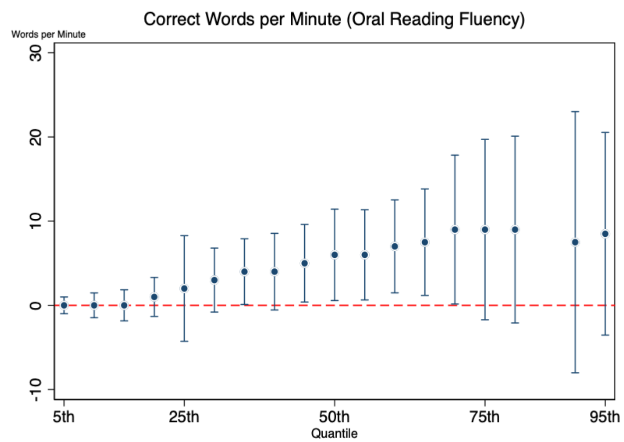
Advanced Skills Distribution



Panel A: Distribution Regression Effects for Oral Reading Fluency



Panel B: Distribution Regression Effects for Reading Comprehension



Panel C: QTEs Effects for Oral Reading Fluency

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

4.2 Mechanisms

To measure how TFLI achieved the large gains we measured changes in teaching quality using a pre-specified index of teaching practices during live lessons, as well as how well schools adhered to TFLI's program. Table 4 presents these results. Column 2 in Panel A shows that overall teaching quality increased by 1.6 SDs relative to the control group and is highly significant. Figure 6 shows the detailed breakdown of specific lesson elements in the index and Figure 7 shows specific teaching practices and student engagement metrics.[22] There are large gains across board for lesson elements. The measurement tool we used for these lesson elements is "opinionated" and looks for specific lesson elements TFLI includes in its scripted lessons, which helps to explain why these effects are as large and consistent as they are. There are large gains for teaching behavior, but they are less precisely estimated than the lesson elements. The largest gains are for moving around the room, supporting struggling learners, and teaching at an appropriate pace. Students show the largest gains in engaging with their workbooks, staying on task, and being familiar with class routines. These gains are consistent with TFLI successfully scripting lessons and scripting reducing wasted time during lessons. Columns 3 and 4 similarly show that compliance with the program increased by 1.3 SDs over the control group. Table B4 shows the detailed breakdown of all the compliance components. The largest increases are in teachers' use of workbooks and teacher guides.[23]

Panels B and C of Table 4 link these large gains in quality to improvements in reading. Panel B shows the raw correlation between teaching quality and overall reading scores within the treatment group. Since our quality measure is opinionated, we would expect to see a significant effect here if e.g. better teachers deviated from the scripted lessons more than poorer teachers (or vice versa). There is no correlation, which is supportive evidence that there is not differential deviation from the scripts based on teacher skill. The same holds for compliance with the program. Panel C shows estimates of the effect of quality of reading

[22] Table 4 shows estimates with the data aggregated at the student level while Table B2 and Table B3 aggregate the data at the teacher level.

[23] Some components of the compliance index exhibited one-sided non-compliance because e.g. control schools could not deliver any of the scripted lessons from TFLI. Table B5 shows estimates for compliance where we set these variables mechanically to zero for control schools.

scores using assignment to treatment as an instrument for teaching quality. This estimates how much a 1-SD in quality or compliance increases test scores, under the assumption that TFLI affects test scores only through each channel. Since both channels move, we know that this exclusion restriction is violated. The point estimates suggest that a 1-SD gain in quality increases test scores by 0.32 SDs, significant at the 0.1 level.

Another avenue through which TFLI may improve test scores is by changing students' at-home behavior. The most notable change with at-home practices is a 9.7% (7.5 pp) increase in students practicing reading at home, although there is no change in how often they do school work at home with their parents/guardians or siblings (Table E3). We also see some changes in student confidence. There is a 27% (3.6 pp) decrease in students who believe they are in the bottom third of their class and a similar 27% (4 pp) reduction for math. The average effect on aspirations is null. There is no change in students' belief they will pass the high school exit exam or get their dream jobs Table E1. Interestingly, Table E4 shows there are also null effects for students' beliefs about the quality of their schooling, although this is likely subject to social desirability bias distorting students' answers.

Table 4
Treatment Effects on Compliance with Program

	Quality		Compliance	
	(1)	(2)	(3)	(4)
	Equiv. Yrs. of Schooling	SDs	Equiv. Yrs. of Schooling	SDs
Panel A: Treatment Effects on School Quality/Compliance				
Treatment Assignment		1.617***		1.296***
S.E.		(0.508)		(0.285)
Effective F (Olea–Pflueger)		10.66		21.21
Adjusted R^2		0.537		0.624
Panel B: Effect on EGRA Scores (OLS)				
Quality/Compliance Index	-0.428	-0.097	0.326	0.074
S.E.	(0.643)	(0.146)	(0.625)	(0.142)
N (# students)	579		552	
C (# stratification cells)	20		19	
Adjusted R^2	0.296		0.240	
Panel C: Effect on EGRA Scores (2SLS)				
Quality/Compliance Index	1.389*	0.316*	1.739*	0.395*
S.E.	(0.750)	(0.170)	(0.968)	(0.220)
N (# students)	1,191		1,119	
C (# schools)	21		20	
Adjusted R^2	0.064		0.142	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Panel B is run only on treated schools. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Figure 6
Lesson Quality Elements

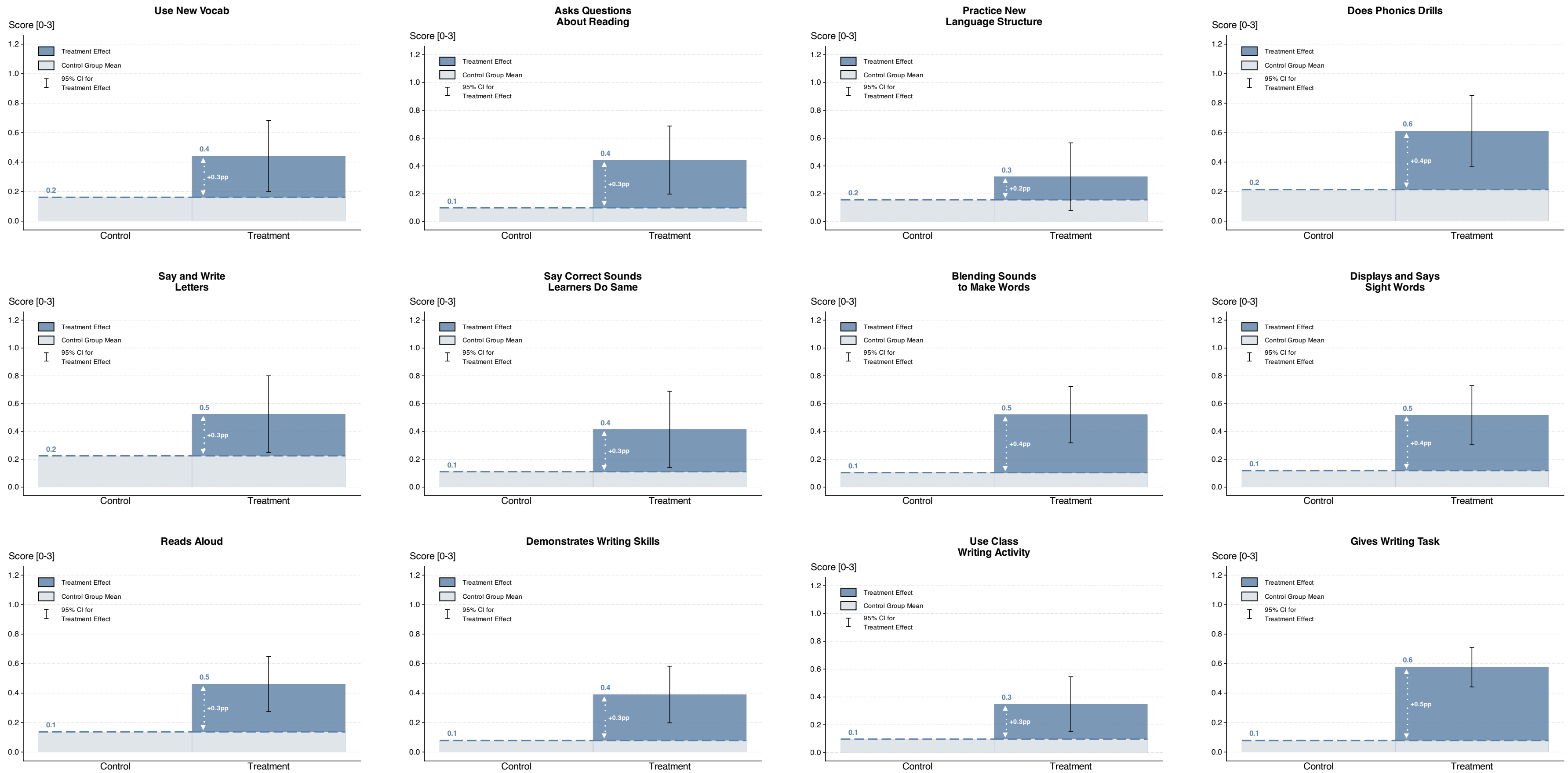
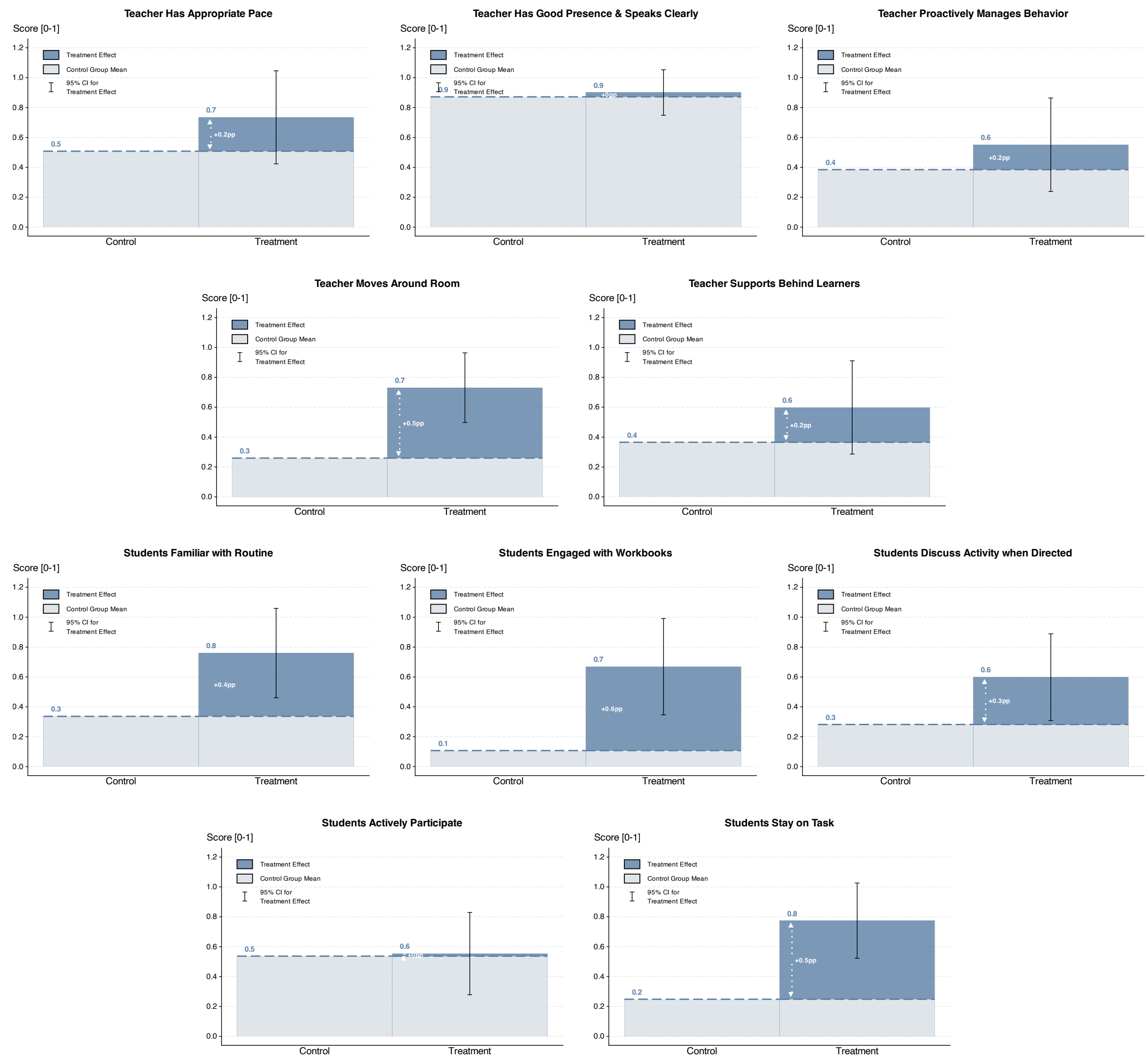


Figure 7

Teacher and Student Quality Elements



4.3 School Leader Training A/B Test

We test these mechanisms further by running an A/B test to see if training school leaders to be better able to coach their teachers can improve teaching quality even more. Since this was an A/B test iterating on aspects of the TFLI program, we only randomized this within the treatment schools. 20 schools received the school leader training, and 20 retained the status-quo TFLI program. Throughout this section, we use a 70% threshold for statistical significance. Lower significance thresholds like this one are standard in A/B testing because the goal is to rapidly iterate on successful improvements, not to collect enough data to cross standard significance thresholds. In accordance with that spirit of rapid testing and iteration, this intervention ran for just two months from the beginning of May through the end of the school year at the end of June 2025.

Table 5 shows the main results of this A/B test. Teaching quality increased by 1.1 SDs relative to the control group, while compliance did not show a significant change. Table C3 shows the detailed breakdown of the lesson components part of the quality index. Unlike the main intervention, there are not across the board improvements in all lesson components. The largest impacts are increases in doing phonics drills, practicing the new language components introduced in a lesson, and doing the writing activity given in the scripted lesson. Although no other components are significant, all are large in magnitude, and about half the size of the quality increase from TFLI overall. Table C4 shows the detailed breakdown of teaching behavior and student engagement. Mostly, these are large but noisily estimated effects. Proactive management of classroom behavior by teachers and participation in-class discussions by students show the largest increases. Table C5 shows that we see no increases in compliance with the program, which was not a main goal of this intervention.

Panel C of Table 5 shows that the learning effects of this intervention are not yet statistically different from zero, but are large in magnitude. Table C1 breaks down the reading index by the individual subtests. These are imprecisely estimated, but larger for more advanced skills. In particular, the effect sizes for oral reading fluency and reading comprehension are 70% and 58% of the main effect size, while the effect size for letter sounds is just 10% of the main treatment effect. This is consistent with school leaders increasing the quality of lessons, which, as we discuss in Section 5, should have larger benefits for more-advanced skills.

Table C2 shows the "long model" (Muralidharan, Romero, and Wüthrich 2025) where we estimate the effect on reading scores with a fully saturated model with dummies for the main treatment and the A/B test treatment. The main pattern of our results is unchanged. There are slightly smaller point estimates for the overall reading index and subtests, but the significance and pattern of skill formation remain unchanged.

Table 5
Quality and Compliance in A/B Test

	Quality		Compliance	
	(1)	(2)	(3)	(4)
	Equiv. Yrs. of Schooling	SDs	Equiv. Yrs. of Schooling	SDs
Panel A: Treatment Effects on School Quality/Compliance				
Treatment Assignment		1.153*		0.317
S.E.		(0.601)		(0.498)
Effective F (Olea–Pflueger)		3.92		0.43
Adjusted R^2		0.224		0.172
Panel B: Effect on EGRA Scores (OLS)				
Quality/Compliance Index	1.274*	0.290*	-0.739	-0.168
S.E.	(0.688)	(0.156)	(1.869)	(0.425)
N (# students)	237		222	
C (# schools)	15		13	
Adjusted R^2	0.181		0.036	
Panel C: Effect on EGRA Scores (2SLS)				
Quality/Compliance Index	0.590	0.134	1.041	0.237
S.E.	(0.955)	(0.217)	(4.684)	(1.065)
N (# students)	562		527	
C (# schools)	35		30	
Adjusted R^2	0.027		-0.033	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Panel B is estimated using only the treatment group. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

5 Theory

In this section, we first review the Cunha and Heckman (2007) model of skill formation. We impose a functional form on the model to study two separate but related skills, which the original setup accommodates, but which is not a focus of their analysis. We use this to explain the broadly-documented pattern in which education interventions have larger impacts on basic skills than on advanced ones. Then we consider a slight modification to the technology that allows teachers to split their time between teaching basic and advanced skills and show that the optimal time allocation decision rule mimics targeted instruction in the vein of TaRL. Finally, we consider the interaction between structured pedagogy and targeted instruction, and provide intuition for why they could be complementary inputs to education.

Our framework builds on models of instruction targeting (Duflo, Dupas, and Kremer 2011) and cumulative learning (Shaikh 2025). Duflo, Dupas, and Kremer show that teachers facing convex payoffs target instruction toward higher-achieving students, and demonstrate empirically that this improves outcomes, but do not model the learning technology that makes targeting beneficial. Shaikh estimates a structural model with cumulative technology where earlier learning increases later productivity, demonstrating dynamic complementarities empirically. We contribute to the theoretical mechanism underlying both findings: basic skills act as prerequisites that constrain advanced learning. This explains why targeted instruction improves outcomes and what creates dynamic complementarities, while generating novel predictions about treatment effect timing and intervention complementarity.

Consider a child who is born with a vector θ_0 of skills. This vector, in principle, contains everything from reading to time management to basketball skills, but here we focus only on literacy skills. In each period t, θ_t denotes the vector of skill stocks. These skills evolve according to the following technology:

$$\theta_{t+1} = f_t(\theta_t, I_t, S_t) \tag{4}$$

Here I_t is a vector of investments in different components of the skills vector θ_t, and S_t is the productivity of time in schools.[24] We can think of targeted instruction as operating through I_t by differentially allocating effort to different skills depending on skill level, and structured pedagogy as operating through S_t which is a general, broad increase in the marginal productivity of time in school. A key feature of the model is that it allows for both dynamic complementarity when $\frac{\partial^2 f_t(\theta_t, I_t, S_t)}{\partial\theta_t \partial I_t} > 0$ (i.e. when the stock of skills accumulated by period $t-1$ makes investment in skills more productive), and self-productivity when $\frac{\partial f_t(\theta_t, I_t, S_t)}{\partial\theta_t} > 0$ (i.e. when skills build on each other to create more skills in the next period). We will focus on the case where self-productivity in advanced skills is constrained by the level of more basic skills.

5.1 Differential Timing of Treatment

For expositional simplicity, consider a student for whom θ has only two skills: basic literacy skills (e.g., letter sound knowledge or phonemic awareness) denoted B_t, and advanced literacy skills (e.g., reading a passage or reading comprehension) denoted A_t. For the moment, we suppress investment in skills, I_t, but will add it back in in the next section for our discussion of targeted instruction. We impose the following functional form on Equation 4:

$$B_{t+1} = B_t + \alpha \cdot S \tag{5}$$

$$A_{t+1} = A_t + \beta \cdot S \cdot h(B_t) \tag{6}$$

$h(\cdot)$ is the "constraint function" which moderates how quickly advanced skills can build up as basic skills hold them back. It has the following properties: 1) $h(0) = 0$ i.e. basic skills are a prerequisite to advanced skills and students cannot develop any advanced skills if they have no basic skills 2) $h'(B_t) > 0$ i.e. the constraint monotonically weakens as students build basic skills 3) $h''(B_t) < 0$ i.e. basic skills eventually stop constraining advanced skill growth, 4) $h(\cdot) \in [0, 1]$.[25]

Consider now a treatment such as structured pedagogy that increases S_t to $S_t + \tau$ in

[24] Cunha and Heckman allow parental characteristics to enter into the skill production function, but we suppress that here as it is not our focus.

[25] Equation 5 is makes the (simplistic) assumption that basic skills build up linearly everywhere. We do this to focus our analysis on inter-skill complementarity.

perpetuity. At time t, treatment effects are given by:

$$\Delta B_t = B_t^T - B_t^C = \alpha\tau t$$
$$\Delta A_t = A_t^T - A_t^C = \underbrace{\beta\tau\sum_{s=0}^{t-1} h(B_s^T)}_{\text{Direct effect}} + \underbrace{\beta S\sum_{s=0}^{t-1}\Big(h(B_s^T) - h(B_s^C)\Big)}_{\text{Indirect effect via } B}$$

While basic skills are affected directly by the treatment, the treatment effect on advanced skills has two distinct components. First, there is the direct effect where τ feeds directly into the stock of advanced skills, moderated by the levels of B_t at each period. The second component is the indirect effect of the treatment. The indirect effect is to increase the stock of basic skills which loosens the constraint function. The treatment thus "unlocks" existing school productivity that students can't harness before they build up a sufficient level of basic skills.

This technology leads us to two predictions for how skills develop.

Prediction 1: *In settings where students have low starting levels of basic skills, treatment effects on basic skills dominate.*

$$\frac{\Delta A_1}{\Delta B_1} = \frac{\beta}{\alpha} \cdot h(B_0) \approx 0$$

Prediction 2: *As treatment continues (or students begin with a large stock of basic skills) treatment effects on advanced skills catch up.*

$$\lim_{t\to\infty} \frac{\Delta A_t}{\Delta B_t} = \frac{\beta}{\alpha}$$

In summary, in the first year of a structured pedagogy-esque education intervention, we predict the largest treatment effects will be for basic phonics skills while advanced passage reading and reading comprehension have smaller effects. As students stay in the program in further years, treatment effects on passage reading and reading comprehension should grow and catch up to the basic skills effects.

This differential treatment effect phenomenon is widely documented in foundational literacy (Piper et al. 2018*d*, Kerwin and Thornton 2021, Fazzio et al. 2021, McManus et al.

2025) and numeracy (Albornoz et al. 2025, McManus et al. 2025), but to our knowledge we provide the first explanation for why skills behave in this way based on human capital theory.

Prediction 1 is borne out in our data. Figure 8 shows that basic phonics skills (letter sounds and initial sound identification) have the largest treatment effects. The most advanced skills are reading a passage and comprehending it; Figure 9 shows that these effects are much smaller than those on basic skills and less precisely estimated. We currently only have one year of data, but in future phases of this study, we will test Prediction 2 to see if advanced skill treatment effects catch up as the intervention continues.

This framework also makes predictions about treatment effect heterogeneity for advanced skills.

Prediction 3: *Treatment effects for advanced skills in the first period are larger for students with higher baseline stocks of basic skills.*

$$\frac{\partial \Delta A_1}{\partial B_0} = \beta\tau \cdot h'(B_0) > 0$$

Since any period can serve as the initial period, this holds for all consecutive periods t and $t+1$ when there is an exogenous treatment after period t. We do not directly test Prediction 3 in the current study since we did not run a baseline survey, but we will have baseline scores for future phases of the study and will test for this then.

5.2 Endogenous Targeted Instruction

Now we add back in time investment by teachers. We will suppress S in this section to focus on the investment dimension. Teachers are endowed with one unit of time and choose what proportion of the time to invest in teaching advanced skills I_t^A, while basic skills get $1 - I_t^A$ units of time. Equations 5 and 6 become the following.

$$B_{t+1} = B_t + \alpha \cdot (1 - I_t^A)$$
$$A_{t+1} = A_t + \beta \cdot I_t^A \cdot h(B_t)$$

For tractability, consider a two period model. The social planner chooses the optimal allocation of time spent teaching advanced skills in periods 1 and 2 to maximize the stock of advanced skills in period 2. Only advanced skills are socially valuable; basic skills' only value is in service of generating advanced skills.

The solution to the social planner's problem mimics the decision rule for targeted instruction. For students with a sufficiently low stock of basic skills, the social planner allocates all instruction time to basic skills. For students with sufficiently high stock of basic skills, they allocate all instruction time to advanced skills. For students between, they split time between basic and advanced skills, and the optimal amount of advanced skill instruction is monotonically increasing in the stock of basic skills. This is summarized in Proposition 1.

Proposition 1 (Optimal Instruction Time Allocation) *Let* $I_1^A \in [0,1]$ *denote the optimal allocation of instruction time to advanced skills in period 1. Define threshold values* $\underline{B}$ *and* $\overline{B}$ *as solutions to:*

$$h(\underline{B}) = \alpha \cdot h'(\underline{B} + \alpha)$$
$$h(\overline{B}) = \alpha \cdot h'(\overline{B})$$

Then the optimal instruction policy is:

$$I_1^A(B_0) = \begin{cases} 0 & \textit{if } B_0 < \underline{B} \quad \textit{(specialize in basics)} \\ \textit{interior solution} & \textit{if } \underline{B} \le B_0 \le \overline{B} \quad \textit{(balanced instruction)} \\ 1 & \textit{if } B_0 > \overline{B} \quad \textit{(specialize in advanced)} \end{cases}$$

For $B_0 \in [\underline{B}, \overline{B}]$*, the interior solution satisfies:*

$$h(B_0) = \alpha \cdot h'(B_0 + \alpha(1 - I_1^A))$$

and is strictly increasing in B_0*:* $\frac{dI_1^A}{dB_0} > 0$*.*

Proof: see Appendix G

In this framework, targeted instruction is more beneficial for certain sets of skills. For skills where the $h'(\cdot)$ function is very steep, advanced skills are heavily constrained by the stock of basic skills. This is the case for highly interrelated skills like letter sound knowledge and non-word reading. For less directly related skills (e.g., general literacy and general math), $h'(\cdot)$ is shallower, so targeted instruction becomes less useful as the skills constrain each other less. This is exactly how targeted instruction is typically deployed: within-subject rather than across subjects.

5.3 Complementarity between targeted instruction and structured pedagogy

Finally, we will add both productivity of time in school and instructional time allocation to the model simultaneously and show that with this functional form, they are complements. Although the complementarity is driven by the multiplicative functional form assumption, the model provides useful intuition for why this relationship may hold. We plan to test explicitly for targeted instruction and structured pedagogy complementarity by randomly varying the intensity of the targeted instruction components of TFLI in future phases of this study.

The full technology we use in this section is given by

$$B_{t+1} = B_t + \alpha \cdot S \cdot (1 - I_t^A)$$
$$A_{t+1} = A_t + \beta \cdot S \cdot I_t^A \cdot h(B_t)$$

Proposition 2 shows that under this functional form assumption for the skill production technology, targeted instruction and structured pedagogy are complementary. Note this is *not* saying investment in advanced skills and structured pedagogy are complementary, but rather that optimally choosing investment levels is complementary with structured pedagogy.

The intuition for Proposition 2 is as follows. When productivity of time at school is low, it makes no difference how well time is allocated: any time spent on anything will not be used well. If a school is poor and students learn nothing, a student who cannot read can equally well learn nothing about complex reading comprehension or simple letter sounds. If productivity at school is high, the converse is true. The higher productivity school time is, the more wasteful it is to assign a remedial student to advanced topics; the student misses out on more valuable time they could be using to build their stock of basic skills.

Proposition 2 (Complementarity of targeted instruction and structured pedagogy) *Define:*

- $V^{TI}(B_0, S) = \max_{I_1^A \in [0,1]} \left\{ \beta S I_1^A h(B_0) + \beta S h(B_0 + \alpha S(1 - I_1^A)) \right\}$

  (value under targeted instruction policy with optimal targeting)

- $V^{uniform}(B_0, S, \bar{I}) = \beta S \bar{I} h(B_0) + \beta S h(B_0 + \alpha S(1 - \bar{I}))$

  (value under uniform policy with fixed $\bar{I}$)

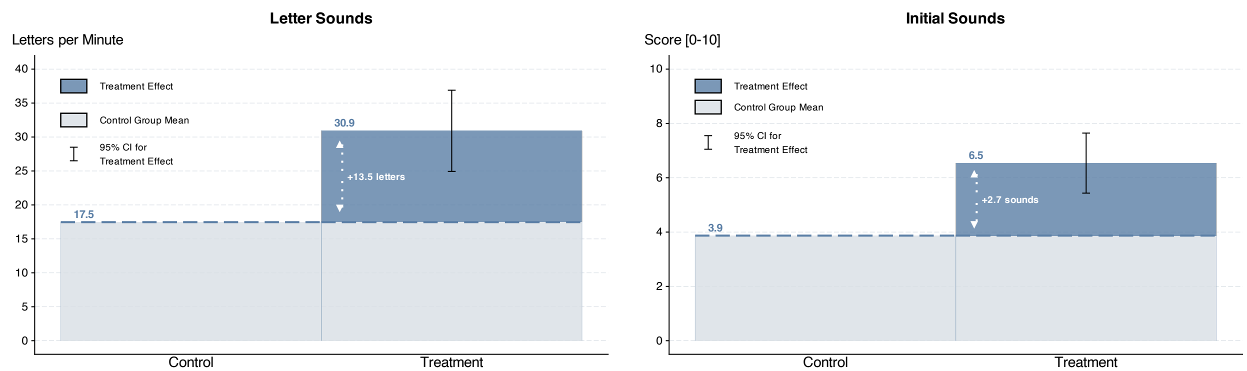
Let $I^*(B_0, S)$ denote the optimal allocation under targeted instruction. Then for any $\bar{I} \neq I^*(B_0, S)$:

$$\frac{\partial V^{TI}(B_0, S)}{\partial S} > \frac{\partial V^{uniform}(B_0, S, \bar{I})}{\partial S}$$

That is, the marginal impact of increasing instructional quality S is strictly larger under a targeted instruction policy that optimally targets instruction than under a uniform policy.

Proof: see Appendix G.

Figure 8
Basic Skills Treatment Effects

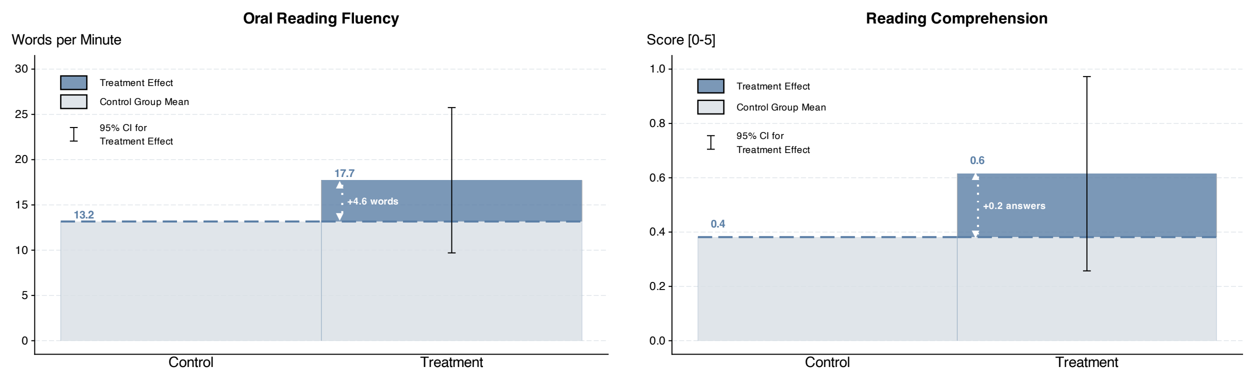


Panel A: Treatment Effects of Letter Sound Knowledge

Panel B: Treatment Effect of Initial Sound Knowledge

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Figure 9
Advanced Skills Treatment Effects



Panel A: Treatment Effects of Letter Sound Knowledge

Panel B: Treatment Effect of Initial Sound Knowledge

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a student-level linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

6 Scalability

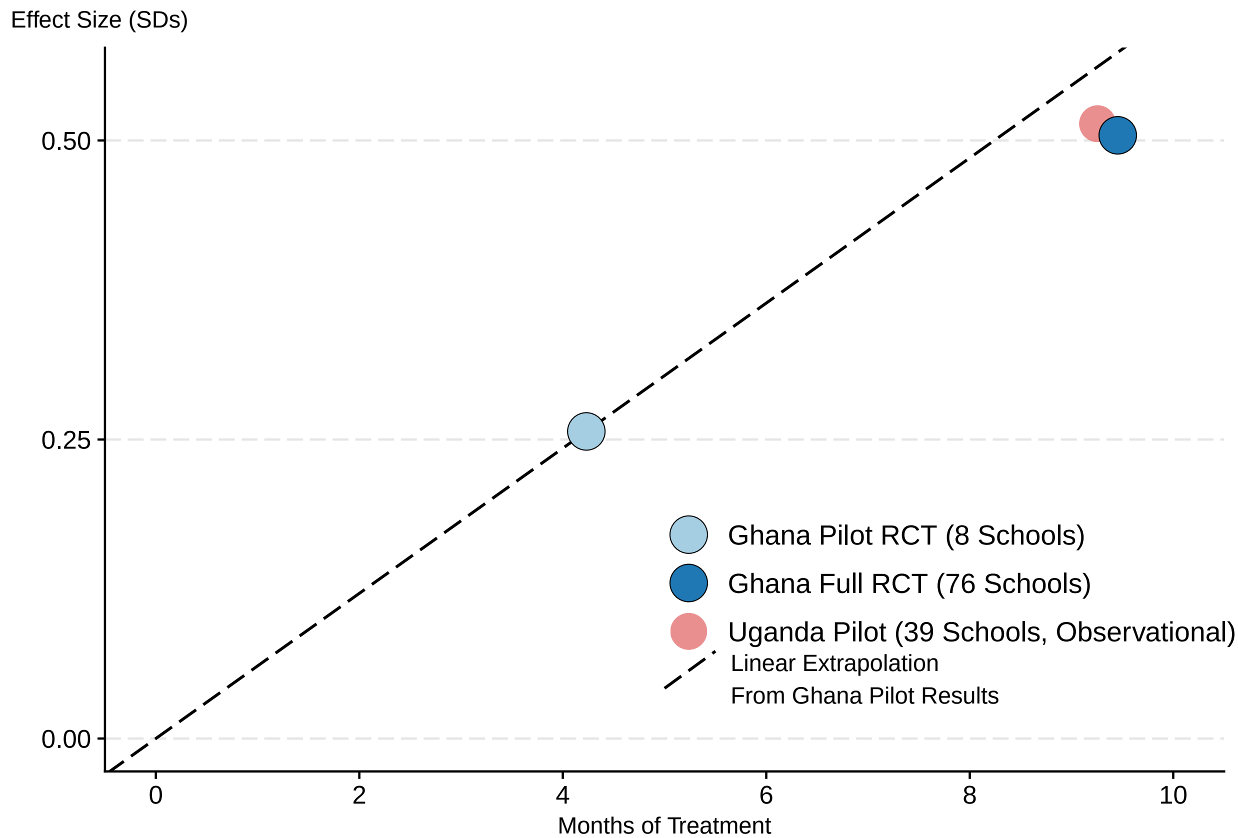
Addressing the learning crisis in Africa will require not just high-impact interventions, but also ones that can be implemented by the existing teaching workforce and rapidly scaled up over school years across time and over education systems across the continent. The former is necessary so that promising annual gains can be capitalized on, by repeating the intervention until the learning gap with rich countries is closed. To fix concepts, assume that the PIRLS (2021) test score gap between South Africa and Singapore (2.99 SDs) is representative of the gap for the entire continent. In that case, TFLI's impacts would need to be repeated nearly six times in order to achieve parity between Africa and rich countries. This is, in principle, possible: our estimates are for a single academic year, and the PIRLS scores are for the end of fourth grade. Since there are two years of kindergarten before B1 in Ghana, a literacy intervention like TFLI could be run six times during primary school in that country.

For this to work, however, the program's impacts would have to scale up over time: doubling the amount of time in the program would need to double treatment effects, or close to it. We show evidence that this may be the case in Figure 10. The dark blue dot in the figure shows our results from the RCT described in this paper on the y-axis and the number of months in the school year on the x-axis. The light blue dot shows the same figures but from a previous pilot RCT that we conducted during the 2024-25 school year, with a sample of just 8 schools (4 treatment and 4 control).[26] The test scores from this pilot RCT were internal Inspiring Teachers exams rather than EGRAs. We apply the exact same analyses to that earlier data as we do in the data from the current study, using as our main outcome a PCA index of all the available subtests from the control-group data. We find a treatment effect of 0.25 SDs. This pilot ran for just over four months, rather than the full nine-month school year. The dashed line extrapolates the gain per month from the pilot RCT to our main study, and finds that we are very close to fitting the linearly-extrapolated trend. This suggests that our gains per month do.

Scaling the program across space is needed so that the intervention can help children not only in the Central Region of Ghana but also across the rest of the country and the rest of

[26] A report with more detailed results on this other experiment is avaialable upon request.

Figure 10
Learning Gains vs. Months of Treatment



the continent. The intervention was designed from the ground up with that in mind, with three pillars supporting its scalability. The first pillar is English-language-first instruction. This is a practical issue, rather than a matter of pedagogical principle or even effectiveness. Mother-tongue-first literacy instruction, wherein students learn to read first in the language they grew up speaking before transitioning to other languages, appears to have benefits (Piper, Zuilkowski, and Ong'ele (2016)) and some highly-effective interventions use this approach (e.g., Kerwin and Thornton 2021).[27] However, there are over 2,000 languages in Africa, and adapting effective teaching materials to all of them would be a massive logistical undertaking. English is an official language or *de facto* lingua franca in 21 of the 58 countries in sub-Saharan Africa, covering 47 percent of the population of the region.[28] A structured

[27] The longer-term evidence on mother-tongue-first instruction is less promising, with some evidence of negative spillovers onto other subjects (Piper et al. 2018*b*).

[28] Estimates from ChatGPT 5.2 Thinking based on a review of Wikipedia and other public sources: `https://chatgpt.com/share/69450567-db40-8010-9fbd-543e1b00143f`

pedagogy approach that begins in English thus has high potential. Moreover, the TFLI intervention achieves large gains in reading in English despite it being the native language of just 15% of our study sample.

Even with an effective English-first literacy intervention, substantial changes will need to be made in order to adapt it to the rest of Africa: countries have major differences in local culture, initial student ability levels, school calendars, national curricula, and more. The TFLI's second and third pillars of scalability make this possible by efficiently leveraging the scarce pool of available teaching experts who can design high-quality literacy lessons. Pillar two is a component-based design approach: lesson plan designers draw on a set of shared building blocks for lessons, and assemble those building blocks in consistent patterns.

Pillar three is the use of generative AI and component-based design to speed up the process of designing new high-quality lesson plans and adapting them across settings. Lesson plan designers use genAI in two key ways. First, they create controlled texts, which help students practice sounds and words they already know and mix in new ones they need to learn. These require following a sets of rules. For example, a story might have to draw from the following list of words students already know, add this new one we are practicing today, match the theme of the lesson, and use existing characters from previous stories. Large language models excel at this task. Guide designers prompt them with the rules, and can focus on evaluating the quality of the texts and on bigger-picture lesson design issues rather than rule-following. The second is illustrating the stories, which can be done far quicker via genAI than by hiring human illustrators (which can take weeks due to multiple rounds of comments and revisions). The use of these tools is also conducive to adapting the lessons across settings: LLMs can quickly draft new versions of lessons that alter key cultural cues and adjust themes and topics to match national curricula, with the lesson designers providing expert supervision rather than focusing on the rote tasks of making these edits.

As a result of this inherent scalability, TFLI is being scaled up both within Ghana and across Africa. The program is operating in 139 schools in Ghana in the 2025-26 school year, including 80 government schools, and Inspiring Teachers has an agreement in place to expand it to 500 schools (400 of them government-run) by 2026-27. They are collaborating with the national and regional offices of the Ghana Education Service to expand the program to every government school in the Central Region by 2029-30 (1,638 in total). This scale-up appears as though it will be highly cost-effective. The 2024-25 version of the intervention had an incremental cost of \$48 per student, and so the cost per 1-SD gain is \$96, which already makes the program competitive with existing interventions. By 2029, Inspiring Teachers' budget model predicts the cost will drop to \$6 per student. If the large impacts of the intervention can be sustained then it will become extremely cost-effective.

This ongoing scale-up also provides evidence that TFLI scales across geography. Inspiring Teachers conducted a pilot of the adaptation of the program to Uganda in the 2025 school year, which ran from February 3 to December 5 in 20 schools in Kanungu District.[29] They selected 19 similar nearby schools being as a comparison group. The treatment assignment was not randomized, and we have a limited set of exogenous covariates to use in our analysis. However, if we construct our outcome in the same way as we specify in Section 3 and condition flexibly on all available control variables, we see an overall test score difference of 0.514 SDs. The detailed results are presented in Table F1; the estimated impacts on individual components differ somewhat from the actual RCT in Ghana. Taken literally, this result has two implications. First, the program scales across space: we see almost the exact same impacts in Uganda as in Ghana. Second, it reinforces our findings on scaling up the program over time: the red dot in Figure 10 shows the test score gain and length of intervention for Uganda, and is also quite close to the linear extrapolation from the 2024-25 Ghana Pilot. The prospects for the future scale-up of the program across Africa are quite promising as well. Inspiring Teachers is already in talks to expand the program to Zambia.

7 Conclusion

Can the learning crisis in Africa be solved? Recent trends have been disheartening. Most efforts to improve education have no hope of closing the colossal gaps between Africa and the world's richest countries. The median education intervention has a causal effect of just 0.1 SDs (Evans and Yuan 2022) while most of Africa is over 3 SDs behind the rich-world educational frontier (PIRLS 2021). There is no realistic prospect of running one of these

[29] School years in Uganda are the same as calendar years.

interventions thirty times, and nobody has tried. This matters not just for the sake of learning itself but for the future of Africa's economy: Engbom et al. (2025) argue that low learning levels in the developing world limit firm size and thus hamstring structural transformation.

We study a program that does make substantial progress toward this goal, accelerating learning by over two years of status quo gains in just one year of intervention: Tools for Foundational Learning Improvement, or TFLI. The program thus joins a handful of programs that have boosted learning by more than 0.5 SDs (e.g., Piper et al. 2018*c* Eble et al. 2021, Fazzio et al. 2021, Gray-Lobe et al. 2022, Buhl-Wiggers et al. 2024). It does so in just a single year of intervention, and in English in a context where fewer than a sixth of students grew up speaking the language, both of which are rare among existing interventions. TFLI achieves this massive progress by capitalizing on strong positive complementarities between structured pedagogy and differentiated instruction. We develop a model of skill formation that shows that these two promising approaches are complementary to one another, and show that it makes predictions that match our results. It also makes forecasts about the patterns that we will observe as we continue to follow the same cohort of children (who will continue to be treated) and as we run additional RCTs to study the program (and can collect data that we currently do not have access to).

A crucial feature of the program is that it can also be scaled. The NGO that created it, Inspiring Teachers, designed it to be adaptable to a wide range of settings, most crucially via the use of generative AI. Lesson plan designers use genAI to rapidly create texts for children to use for reading practice that fit the needs of the lesson in question, and illustrations to accompany the stories in the lessons. This approach allows for faster curriculum alignment and has already paid dividends, with a successful pilot-test in Uganda. The future scale-up prospects of the program look bright: the organization has laid the groundwork to expand to all government schools in Ghana's Central Region and also into Zambia. And the intervention's impacts appear to scale almost linearly with time in the program. Our results thus suggest that with the right programs, the staggering learning gaps between Africa and developed countries can, in fact, be remedied.

References

Aizer, Anna, and Flávio Cunha. 2012. "The Production of Human Capital: Endowments, Investments and Fertility." National Bureau of Economic Research NBER Working Paper 18429.

Akaguri, Luke. 2014. "Fee-Free Public or Low-Fee Private Basic Education in Rural Ghana: How Does the Cost Influence the Choice of the Poor?" *Compare: A Journal of Comparative and International Education*, 44(2): 140–161.

Albornoz, Facundo, Gonzalo Almeyda Torres, María Lombardi, Victoria Oubiña, and Pablo Zoido Lobaton. 2025. "Remote Tutoring in Latin America." *Journal of Development Economics*, 103687.

Alvarez Marinelli, Horacio, Izzy Boggild-Jones, Michael Crawford, Margaret Mary Dubeck, Dhir Jhingran, Christopher Joseph Lack, Nompumelelo Mohohlwane, Maria Eugenia Oviedo Buitrago, Benjamin Piper, and Jaime Saavedra. 2025. "Effective Reading Instruction in Low- and Middle-Income Countries: What the Evidence Shows." GEEAP. Publisher: World Bank Group.

Angrist, Joshua, Victor Chernozhukov, and Iván Fernández-Val. 2006. "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure." *Econometrica*, 74(2): 539–563. Publisher: Econometric Society.

Angrist, Noam, and Rachael Meager. 2023. "Implementation Matters: Generalizing Treatment Effects in Education." *SSRN Electronic Journal*.

Angrist, Noam, Claire Cullen, and Janica Magat. 2025. "Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale." Working Paper. Publisher: What Works Hub for Global Education.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives*, 31(4): 73–102.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics.*

Beg, Sabrin A., Anne E. Fitzpatrick, and Adrienne Lucas. 2023. "Managing to Learn."

Brion, Casey. 2020. "Low-Fee Private Schools in Ghana: A Mixed-Methods Study of English as a Marketable Asset." ERIC full-text PDF.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Sebastián Muñoz, Jeffrey A. Smith, and Rebecca L. Thornton. 2024. "Some Children Left Behind: Variation in the Effects of an Educational Intervention." *Journal of Econometrics*, forthcoming.

Carneiro, Pedro, Yyannú Cruz-Aguayo, Rafael Hernández-Pachón, and Norbert Schady. 2025. "Dynamic Complementarity in Elementary Schools: Experimental Estimates from Ecuador." Working Paper.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions - Chernozhukov - 2013 - Econometrica - Wiley Online Library."

Correia, Sergio. 2016. "Estimating Multi–Way Fixed Effect Models with reghdfe."

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review*, 97(2): 31–47.

Curto, Stefano, and Michelle C. Keane. 2025. "Transforming Ghana in a Generation: 2025 Policy Notes." World Bank, Washington, DC.

Day Ashley, Laura, Claire Mcloughlin, Monazza Aslam, Jakob Engel, Jenny Wales, Shailen Rawal, Richard Batley, Geeta Kingdon, Susan Nicolai, and Pauline Rose. 2014. "The Role and Impact of Private Schools in Developing Countries: A Rigorous Review of the Evidence." Department for International Development (DFID), London, UK.

de Chaisemartin, Clément, and Jaime Ramirez-Cuellar. 2024. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics*, 16(1): 193–212.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5): 1739–1774.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron, Stijn Vansteelandt, Peter Boone, and Diana Elbourne. 2021. "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia." *Journal of Development Economics*, 148: 102539.

Engbom, Niklas, Hannes Malmberg, Tommaso Porzio, Federico Rossi, and Todd Schoellman. 2025. "Economic Development According to Chandler."

Evans, David, and Fei Yuan. 2022. "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis*, forthcoming.

Evans, David K, and Fei Yuan. 2019. "Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms." World Bank Working Paper 8752.

Fazzio, Ila, Alex Eble, Robin L. Lumsdaine, Peter Boone, Baboucarr Bouy, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Simon Johnson, and Ana Filipa Silva. 2021. "Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau." *Journal of Public Economics*, 199: 104385.

Ghana Statistical Service. 2019. "Ghana Living Standards Survey (GLSS 7): Main Report." Ghana Statistical Service, Accra, Ghana. Often cited as 2017–2018/"GLSS7".

Gilraine, Michael. 2017. "School Accountability and the Dynamics of Human Capital Formation." Working paper, University of Toronto (often cited as 2018 working paper).

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W. Ozier. 2022. "Can Education be Standardized? Evidence from Kenya." SSRN Scholarly Paper 4129184, Rochester, NY.

Ibrahim, Hosam, Andreas de Barros, Sarah Deschênes, and Paul Glewwe. 2024. "Prospective Evidence on Successful Remediation in Morocco's Public Primary Schools." Working Paper.

Kerwin, Jason, Nada Rostom, and Olivier Sterck. 2025. "Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix It."

Kerwin, Jason T., and Rebecca L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *The Review of Economics and Statistics*, 103(2): 251–264.

Koenker, Roger, and Zhijie Xiao. 2003. "Inference on the Quantile Regression Process - Koenker - 2002 - Econometrica - Wiley Online Library."

Kook, Lucas, and Niklas Pfister. 2025. "Instrumental variable estimation of distributional causal effects." *Electronic Journal of Statistics*, 19(2): 5249–5288. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Lee, David S. 2005. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects."

List, John A., Jeffrey A. Livingston, and Susanne Neckermann. 2011. "Harnessing Complementarities in the Educational Production Function." Working Paper.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania*." *The Quarterly Journal of Economics*, 134(3): 1627–1673.

McKenzie, David. 2025. "Designing and Analysing Powerful Experiments: Practical Tips for Applied Researchers." *Fiscal Studies*, 46(3): 305–322.

McManus, Jeffery, Mico Rudasingwa, Ewoud Nijhof, Kenna Mokobi, Serigne Fallou Deme, James Kiawoin, Felipe Acero Garay, Leah Mwai, and Cassandre Pignon. 2025. "Improving learning outcomes for out-of-school children: evidence from a randomized evaluation of an accelerated learning program in Liberia." *Education Economics*, 33(1): 19–38.

Ministry of Education (Ghana). 2018. "Education Strategic Plan 2018–2030." Ministry of Education, Republic of Ghana, Accra, Ghana.

Molina, Ezequiel, Adelle Pushparatnam, Rimm-Kaufman Sara Elisabeth, and Keri Ka-Yee Wong. 2018. "Evidence-Based Teaching: Effective Teaching Practices in Primary School Classrooms."

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*, 109(4): 1426–1460.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich. 2025. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments." *The Review of Economics and Statistics*, 107(3): 589–604.

Piper, Benjamin, Joseph Destefano, Esther M. Kinyanjui, and Salome Ong'ele. 2018*a*. "Scaling up Successfully: Lessons from Kenya's Tusome National Literacy Program." *Journal of Educational Change*, 19(3): 293–321.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Dunston Kwayumba, and Arbogast Oyanga. 2018*b*. "Examining the Secondary Effects of Mother-Tongue Literacy Instruction in Kenya: Impacts on Student Learning in English, Kiswahili, and Mathematics." *International Journal of Educational Development*, 59: 110–127.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei, and Simon J. King. 2018*c*. "Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides." *World Development*, 106: 324–336.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei, and Simon J. King. 2018*d*. "Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides." *World Development*, 106: 324–336.

Piper, Benjamin, Stephanie S. Zuilkowski, and Salome Ong'ele. 2016. "Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya." *Comparative Education Review*, 60(4): 776–807.

PIRLS. 2021. "Results – Countries' Reading Achievement – PIRLS 2021 – PIRLS 2021."

Rudalevige. 2003. "The Politics of No Child Left Behind." *Education Next*, 3(4).

Shaikh, Hammad. 2025. "Identifying a Cumulative Learning Technology: Evidence from Online Learning." Working Paper.

Social Impact, Inc. 2018. "Ghana Early Grade Reading Program Impact Evaluation: 2017 Baseline Report." United States Agency for International Development (USAID), Washington, DC. Prepared independently by Social Impact, Inc.

Todd, Petra E., and Kenneth I. Wolpin. 2007. "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps." *Journal of Human Capital*, 1(1): 91–136.

UNESCO. 2023. "Ghana: Spotlight on SDG 4 (Advocacy Brief)." UNESCO Global Education Monitoring Report.

UNICEF, and Ministry of Education (Ghana). 2023. "Data Must Speak: How Teacher and Headteacher Characteristics and Behaviors Relate to Student Learning in Ghana." UNICEF Office of Research – Innocenti Policy Brief.

World Bank Group. 2018. "World Development Report 2018: Learning to Realize Education's Promise."

A Appendix Tables and Figures

Table A1
Attrition Patterns by Demographics

	(1)	(2)	(3)	(4)
Treat	0.031	0.010	-0.013	-0.033
S.E.	(0.037)	(0.031)	(0.009)	(0.033)
Male × Treat				0.036
S.E.				(0.032)
Age 5 × Treat				-0.016
S.E.				(0.011)
Age 6 × Treat				-0.011
S.E.				(0.010)
Age 7 × Treat				-0.010
S.E.				(0.010)
Age 8 × Treat				-0.013
S.E.				(0.010)
Age 9 × Treat				-0.040*
S.E.				(0.020)
Age 10 × Treat				-0.008
S.E.				(0.013)
Male × Age 5 × Treat				0.007
S.E.				(0.013)
Male × Age 6 × Treat				0.006
S.E.				(0.008)
Male × Age 7 × Treat				0.009
S.E.				(0.009)
Male × Age 8 × Treat				0.011
S.E.				(0.011)
Male × Age 9 × Treat				0.043***
S.E.				(0.016)
Missing Gender × Age 6 × Treat				-0.498***
S.E.				(0.169)
Missing Gender × Age 8 × Treat				-0.714***
S.E.				(0.243)
Stratification FE	No	Yes	Yes	Yes
Baseline controls	No	No	Yes	Yes
Observations	1,643	1,643	1,643	1,643
Clusters	21	21	21	21

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A2
Post Attrition Balance Table – Students

	(1) Control Mean (SD)	(2) Treatment Mean (SD)	(3) Reg. Adj. Diff (T-C) (p-value)	(4) Obs.
Male	0.550	0.536	-0.009	1,293
	(0.498)	(0.499)	(0.763)	
Student Age (Years)	6.558	6.448	-0.104	1,280
	(0.833)	(0.757)	(0.348)	
Student's Perception of Mother's Age	0.358	0.389	0.023	1,322
	(0.480)	(0.488)	(0.622)	
Answered Mother's Age	37.796	30.972	-4.855	492
	(126.115)	(20.938)	(0.404)	
Student's Perception of Father's Age	0.330	0.360	0.024	1,322
	(0.470)	(0.480)	(0.395)	
Answered Father Age	33.481	37.406	3.671	454
	(23.289)	(26.622)	(0.385)	
Sibling at School	0.691	0.737	0.051	1,298
	(0.462)	(0.440)	(0.233)	
English Spoken with Teacher	0.779	0.801	0.025	1,274
	(0.415)	(0.400)	(0.696)	
English Spoken with Friends	0.642	0.620	-0.020	1,274
	(0.480)	(0.486)	(0.768)	
English Spoken at Home	0.139	0.153	0.021	1,279
	(0.347)	(0.360)	(0.386)	
Fante Spoken with Teacher	0.180	0.152	-0.054	1,274
	(0.385)	(0.359)	(0.187)	
Fante Spoken with Friends	0.273	0.295	-0.020	1,274
	(0.446)	(0.456)	(0.696)	
Fante Spoken at Home	0.614	0.685	-0.023	1,279
	(0.487)	(0.465)	(0.531)	
Twi Spoken with Teacher	0.043	0.047	0.027	1,274
	(0.202)	(0.212)	(0.500)	
Twi Spoken with Friends	0.001	0.000	-0.001	1,274
	(0.038)	(0.000)	(0.348)	
Twi Spoken at Home	0.237	0.153	0.002	1,279
	(0.425)	(0.360)	(0.947)	
Family Owns a TV	0.879	0.891	0.007	1,298
	(0.327)	(0.312)	(0.760)	
Family Owns a Refridgerator	0.694	0.761	0.061	1,298
	(0.461)	(0.427)	(0.233)	
Family Owns a Car	0.380	0.374	-0.015	1,296
	(0.486)	(0.484)	(0.743)	
Family Owns an Oven	0.612	0.679	0.073	1,291
	(0.488)	(0.467)	(0.349)	
Family Owns a Bicycle	0.374	0.411	0.040	1,296
	(0.484)	(0.492)	(0.114)	
Family Owns a SmartPhone	0.833	0.838	0.013	1,293
	(0.374)	(0.368)	(0.636)	
Joint F-stat			2.50	
RI p-value			0.19	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Joint F-stat based on Kerwin, Rostom, and Sterck (2025). Estimated only on students surveyed at endline. Differences in column 3 are estimated using a linear regression that controls for stratification cell indicators. Heteroskedasticity-robust p-values, clustered by stratification cell. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A3
Post Attrition Balance Table – Teachers

	(1) Control Mean (SD)	(2) Treatment Mean (SD)	(3) Reg. Adj. Diff (T-C) (p-value)	(4) Obs.
Male	0.171	0.206	0.021	69
	(0.382)	(0.410)	(0.848)	
Teacher Age (Years)	25.562	27.429	2.098	67
	(9.259)	(11.173)	(0.510)	
Highest Qualification: Bachelor's Degree	0.121	0.114	0.011	68
	(0.331)	(0.323)	(0.911)	
Highest Qualification: Certificate	0.121	0.086	-0.042	68
	(0.331)	(0.284)	(0.657)	
Highest Qualification: SHS	0.697	0.800	0.084	68
	(0.467)	(0.406)	(0.470)	
Highest Qualification: No SHS	0.061	0.000	-0.053	68
	(0.242)	(0.000)	(0.251)	
Years of Teaching Experience	5.914	7.026	0.606	69
	(6.693)	(7.812)	(0.727)	
Years at Current School	4.282	5.864	1.907	71
	(6.103)	(7.400)	(0.365)	
Has Functional Phone	0.706	0.694	0.000	70
	(0.462)	(0.467)	(1.000)	
Joint F-stat			0.40	
RI p-value			0.91	

Notes: Sample is all grade one teachers at the 80 study schools who were present for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Joint F-stat based on Kerwin, Rostom, and Sterck (2025). Estimated only on teachers surveyed at endline. Differences in column 3 are estimated using a linear regression that controls for stratification cell indicators. Heteroskedasticity-robust p-values, clustered by stratification cell. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A4

Post Attrition Balance Table – School Leaders

	(1) Control Mean (SD)	(2) Treatment Mean (SD)	(3) Reg. Adj. Diff (T-C) (p-value)	(4) Obs.
Male	0.706	0.788	0.128	67
	(0.462)	(0.415)	(0.268)	
School Leader Age (Years)	35.818	40.125	6.580	65
	(13.873)	(24.005)	(0.273)	
Highest Qualification: Master's Degree	0.000	0.031	0.037	64
	(0.000)	(0.177)	(0.433)	
Highest Qualification: Bachelor's Degree	0.375	0.469	0.134	64
	(0.492)	(0.507)	(0.341)	
Highest Qualification: Certificate	0.250	0.219	-0.024	64
	(0.440)	(0.420)	(0.865)	
Highest Qualification: SHS	0.344	0.281	-0.122	64
	(0.483)	(0.457)	(0.365)	
Highest Qualification: No SHS	0.031	0.000	-0.024	64
	(0.177)	(0.000)	(0.427)	
Years of Teaching Experience	14.348	16.107	2.605	66
	(9.374)	(13.321)	(0.600)	
Years as School Leader	9.591	10.811	1.285	67
	(7.307)	(10.339)	(0.676)	
Years as Leader at Current School	6.946	6.710	-1.570	67
	(6.990)	(8.093)	(0.519)	
Has Functional Phone	0.121	0.061	0.012	66
	(0.331)	(0.242)	(0.890)	
Joint F-stat			0.69	
RI p-value			0.78	

Notes: Sample is all school leaders at the 80 study schools who were present for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Joint F-stat based on Kerwin, Rostom, and Sterck (2025). Estimated only on school leaders surveyed at endline. Differences in column 3 are estimated using a linear regression that controls for stratification cell indicators. Heteroskedasticity-robust p-values, clustered by stratification cell. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A5

Post Attrition Balance Table – Schools

	(1) Control Mean (SD)	(2) Treatment Mean (SD)	(3) Reg. Adj. Diff (T-C) (p-value)	(4) Obs.
Proportion Male	0.560	0.536	-0.015	76
	(0.138)	(0.202)	(0.714)	
Number of Students in Sample	18.289	15.395	-3.255	76
	(11.943)	(13.312)	(0.170)	
School Fee (GHS)	184.222	195.946	11.625	73
	(94.612)	(95.908)	(0.616)	
KG1 Boys	10.763	8.737	-2.558*	67
	(7.995)	(7.989)	(0.093)	
KG1 Girls	10.711	8.053	-2.977*	67
	(7.832)	(7.559)	(0.079)	
KG2 Boys	10.474	8.105	-2.535	67
	(7.982)	(7.435)	(0.113)	
KG2 Girls	9.447	7.842	-1.686	67
	(7.493)	(7.250)	(0.278)	
Basic 1 Boys	11.342	9.474	-1.267	67
	(9.151)	(9.369)	(0.510)	
Basic 1 Girls	9.711	9.184	-0.140	67
	(7.819)	(8.791)	(0.931)	
Basic 2 Boys	10.132	8.895	-0.523	67
	(7.864)	(7.468)	(0.730)	
Basic 2 Girls	10.842	8.474	-2.209	67
	(8.251)	(7.849)	(0.202)	
Basic 3 Boys	9.342	9.500	1.070	67
	(8.218)	(11.640)	(0.622)	
Basic 3 Girls	8.947	8.711	-0.151	67
	(7.669)	(9.918)	(0.931)	
Basic 4 Boys	7.711	8.211	1.221	67
	(6.932)	(9.743)	(0.535)	
Basic 4 Girls	8.974	8.342	-0.174	67
	(8.707)	(9.490)	(0.931)	
Basic 5 Boys	6.447	7.868	3.081*	67
	(6.246)	(10.044)	(0.094)	
Basic 5 Girls	7.447	8.184	1.523	67
	(6.717)	(10.379)	(0.449)	
Basic 6 Boys	5.974	7.289	2.849	67
	(5.819)	(10.449)	(0.174)	
Basic 6 Girls	6.184	7.658	2.570	67
	(6.120)	(10.103)	(0.183)	
JHS 1 Boys	7.658	6.132	-1.233	67
	(19.495)	(8.537)	(0.778)	
JHS 1 Girls	4.421	5.132	1.560	66
	(4.694)	(7.936)	(0.316)	
JHS 2 Boys	3.816	5.605	2.837*	67
	(4.398)	(8.156)	(0.095)	
JHS 2 Girls	3.868	4.447	1.337	67
	(3.807)	(6.833)	(0.316)	
JHS 3 Boys	2.342	3.605	2.035*	67
	(3.843)	(5.475)	(0.070)	
JHS 3 Girls	2.500	3.737	1.767	67
	(4.285)	(6.079)	(0.154)	
Joint F-stat			1.33	
RI p-value			0.24	

Notes: Sample is 80 study schools selected at the beginning of the 2024-25 academic year. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Joint F-stat based on Kerwin, Rostom, and Sterck (2025). Estimated only on schools included in endline. Differences in column 3 are estimated using a linear regression that controls for stratification cell indicators. Heteroskedasticity-robust p-values, clustered by stratification cell. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A6
Treatment Robustness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall Reading PCA Index (SDs)							
Treatment Effect	0.510	0.503	0.516	0.515	0.504	0.436	0.505	0.437
S.E.	(0.205)	(0.208)	(0.203)	(0.207)	(0.215)	(0.217)	(0.212)	(0.214)
R.I. *p*-value	[0.007]***	[0.012]**	[0.005]***	[0.012]**	[0.013]**	[0.031]**	[0.005]***	[0.035]**
N (# students)	1,298	1,298	1,298	1,298	1,298	1,298	1,298	1,298
C (# stratification cells)	21	21	21	21	21	21	20	20
Adjusted R^2	0.054	0.052	0.072	0.070	0.175	0.161	0.176	0.162
Control mean	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
Control SD	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Treatment Variable	Assigned	Received	Assigned	Received	Assigned	Received	Assigned	Received
Age-Sex Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Strat. Cell FE	None	None	None	None	Original	Original	Pooled†	Pooled†

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. In odd columns overall reading PCA index is normalized with respect to assigned control group. In even columns overall reading PCA index is normalized with respect to received control group. † indicates singleton cells combined (cell 14 merged into cell 7). Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). Randomization inference *p*-values, clustered by school and stratified by stratification cell, in brackets: *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A7

Lee Bounds for Reading Outcomes

	(1) Overall Reading PCA Index	(2) Listening Comprehension	(3) Letter Names	(4) Letter Sounds	(5) Initial Sound Identification	(6) Familiar Word Reading	(7) Non-word Reading	(8) Oral Reading Fluency	(9) Reading Comprehension
	Equiv. Yrs. of Schooling	Score [0-5]	CLPM	CLPM	Score [0-10]	CWPM	CWPM	CWPM	Score [0-5]
Upper Bound	0.560**	0.188	4.230	14.507***	2.886***	4.648*	4.292***	5.422	0.257
S.E.	(0.225)	(0.146)	(3.614)	(3.391)	(0.646)	(2.723)	(1.331)	(4.036)	(0.182)
Lower Bound	0.394**	0.093	1.435	11.897***	2.508***	2.626	3.022**	2.138	0.091
S.E.	(0.196)	(0.136)	(2.796)	(2.872)	(0.503)	(2.584)	(1.216)	(3.772)	(0.184)
Control-group values									
Mean	-0.000	0.540	32.660	17.459	3.870	8.331	2.974	13.169	0.381
SD	1.000	0.920	19.980	17.496	3.760	13.668	7.441	21.426	0.926

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table A8
PCA Weights

Variable	(1) Absolute Weight	(2) Relative Weight
Correct Answers Listening	0.093	0.745
Correct Letters Names Per Minute	0.095	0.763
Correct Letter Sounds Per Minute	0.129	1.035
Initial Sounds	0.115	0.923
Correct Familiar Words Per Minute	0.149	1.189
Correct Non-Words Per Minute	0.132	1.058
Correct Words Per Minute	0.150	1.203
Correct Answers Reading	0.136	1.085

Notes: Score is the weighted average of the subtest scores, where the weights are the first principal component of the control-group data across all English EGRA components we tested in this wave of data collection, for every student in the sample. We standardize each subtest score by the control-group mean and SD before running PCA. Column 1 shows the raw weights given to each component. Column 2 shows the weights rescaled to have a mean of 1, or the relative weight given to each component.

B Detailed Quality & Compliance

Table B1

Treatment Effects on Compliance with Program (Control Schools Mechanically set to 0)

	Quality		Compliance		Compliance (Mech. 0s)	
	(1)	(2)	(3)	(4)	(5)	(6)
	Equiv. Yrs. of Schooling	SDs	Equiv. Yrs. of Schooling	SDs	Equiv. Yrs. of Schooling	SDs
Panel A: Treatment Effects on School Quality/Compliance						
Treatment Assignment		1.617		0.689		2.851***
S.E.		(0.508)		(0.207)		(0.397)
Effective F (Olea–Pflueger)		10.66		21.21		126.33
Adjusted R^2		0.537		0.967		0.989
Panel B: Effect on EGRA Scores (OLS)						
Quality/Compliance Index	-0.428	-0.097	0.326	0.074	1.561	0.355
S.E.	(0.643)	(0.146)	(0.625)	(0.142)	(0.000)	(0.000)
N (# students)	579		552		421	
C (# stratification cells)	20		19		19	
Adjusted R^2	0.296		0.240		0.279	
Panel C: Effect on EGRA Scores (2SLS)						
Quality/Compliance Index	1.389	0.316	1.739	0.395	0.647*	0.147*
S.E.	(0.750)	(0.170)	(0.968)	(0.220)	(0.383)	(0.087)
N (# students)	1,191		1,119		988	
C (# stratification cells)	21		20		20	
Adjusted R^2	0.064		0.142		0.188	

Notes: All regressions are aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. In columns (5) and (6) components of the compliance index are mechanically set to 0 for treatment schools since they are part of the intervention. Panel B is estimated using only treatment schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table B2

Treatment Effects on Lesson Quality

		Oral Language			Phonics			Reading			Writing		
	(1) Quality Index	(2) Use New Vocab	(3) Asks Questions About Reading	(4) Practice New Language Structure	(5) Does Phonics Drills	(6) Say and Write Letters	(7) Say Correct Sounds Learners do Same	(8) Blending Sounds to Make Words	(9) Displays and Say Sight Words	(10) Read Aloud	(11) Use Writing Examples	(12) Use Class Writing Activity	(13) Gives Writing Task
Panel A: Treatment Effects on School Quality													
Treatment Assignment	1.891***	0.331**	0.307**	0.232	0.455***	0.379***	0.382**	0.359**	0.339**	0.298**	0.362***	0.322***	0.560***
S.E.	(0.475)	(0.140)	(0.120)	(0.137)	(0.129)	(0.121)	(0.147)	(0.130)	(0.127)	(0.129)	(0.099)	(0.099)	(0.089)
Adjusted R^2	0.393	0.211	0.164	0.284	0.346	0.369	0.298	0.278	0.207	0.130	0.189	0.063	0.429
Panel B: Effect on EGRA Scores (OLS)													
Quality Index	-0.249	-0.776	-0.413	-0.604	0.104	0.013	-0.000	-0.119	-0.498	0.685	-0.175	-0.852	-0.677
S.E.	(0.291)	(0.464)	(0.860)	(0.681)	(1.157)	(1.138)	(1.055)	(0.714)	(0.602)	(0.440)	(0.823)	(0.637)	(0.545)
N (# of teachers)	34	34	34	34	34	34	34	34	34	34	34	34	34
C (# of stratification cells)	20	20	20	20	20	20	20	20	20	20	20	20	20
Adjusted R^2	0.945	0.954	0.934	0.943	0.928	0.928	0.928	0.928	0.942	0.950	0.929	0.955	0.953

Notes: In columns 2-13, units are a 0-3 rating scale. Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table B3

Treatment Effects on Lesson Quality (Cont'd)

	Teacher						Student				
	(1) Quality Index	(2) Appropriate Pace	(3) Good Presence & Speaks Clearly	(4) Proactively Manages Behavior	(5) Move Around Room	(6) Support Behind Learners	(7) Familiar with Routines	(8) Engaged with Workbooks	(9) Discuss After Teacher Direction	(10) Actively Involved in Activities	(11) Stay on Task
Panel A: Treatment Effects on School Quality											
Treatment Assignment	1.891***	0.243	0.037	0.208	0.576***	0.281	0.470**	0.706***	0.246	-0.002	0.406**
S.E.	(0.475)	(0.167)	(0.104)	(0.164)	(0.120)	(0.192)	(0.171)	(0.123)	(0.168)	(0.159)	(0.164)
Adjusted R^2	0.393	0.167	0.285	-0.075	0.468	0.084	0.159	0.439	0.072	0.184	0.270
Panel B: Effect on EGRA Scores (OLS)											
Quality Index	-0.249	0.401	-0.136	-0.147	-0.444	-0.523	0.560	0.416	-0.752***	-0.363	-0.036
S.E.	(0.291)	(0.568)	(0.383)	(0.519)	(0.474)	(0.375)	(0.500)	(0.435)	(0.114)	(0.464)	(0.822)
N (# of teachers)	34	34	34	34	34	34	34	34	34	34	34
C (# of stratification cells)	20	20	20	20	20	20	20	20	20	20	20
Adjusted R^2	0.945	0.935	0.928	0.930	0.939	0.949	0.944	0.939	0.964	0.940	0.928

Notes: In columns 2-13, units are a 0-3 rating scale. Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table B4

Treatment Effects on Program Compliance

	Teacher			Student	
	(1) Compliance Index	(2) Used Workbooks (%)	(3) Used Teacher Guide (%)	(4) Prop One - One Assessments	(5) Used Report Cards (%)
Panel A: Treatment Effects on Compliance					
Treatment Assignment	1.520***	0.727***	0.712***	0.256	0.043
S.E.	(0.470)	(0.105)	(0.105)	(0.395)	(0.065)
Effective F (Olea–Pflueger)	21.21	68.25	24.41	0.46	0.11
Adjusted R^2	0.314	0.631	0.565	-0.005	-0.025
Panel B: Effect on EGRA Scores (OLS)					
Compliance Index	0.166	0.551	0.182	0.146	-0.059
S.E.	(0.224)	(0.386)	(0.521)	(0.213)	(1.167)
N (# of teachers)	30	34	34	33	35
C (# of stratification cells)	19	20	20	20	21
Adjusted R^2	0.944	0.935	0.929	0.946	0.934
Panel C: Effect on EGRA Scores (2SLS)					
Compliance Index	0.395*	0.732**	0.863**	3.355	29.935
S.E.	(0.220)	(0.329)	(0.435)	(5.352)	(86.681)
N (# students)	1,119	1,191	1,191	1,221	1,238
C (# stratification cells)	20	21	21	21	21
Adjusted R^2	0.142	0.155	0.148	-4.599	-23.168
Control-group values					
Mean	-0.000	0.129	0.097	3.062	1.909
SD	1.000	0.341	0.301	0.878	0.292

Notes: Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table B5

Treatment Effects on Program Compliance

	Teacher						Student	
	(1) Compliance Index	(2) Prop. ORF Assessments Run (%)	(3) Prop. of Lessons Delivered (%)	(4) Prop. of PLC Meetings Held (%)	(5) Used Workbooks (%)	(6) Used Teacher Guide (%)	(7) Prop One - One Assessments	(8) Used Report Cards (%)
Panel A: Treatment Effects on Compliance								
Treatment Assignment	2.851***	0.337***	0.850***	0.336***	0.735***	0.531***	-0.749***	-0.107
S.E.	(0.397)	(0.056)	(0.036)	(0.077)	(0.121)	(0.084)	(0.204)	(0.072)
Effective F (Olea–Pflueger)	126.33	346.01	1046.14	19.28	68.25	24.41	0.46	0.11
Adjusted R^2	0.989	0.876	0.990	0.839	0.971	0.943	0.893	0.766
Panel B: Effect on EGRA Scores (OLS)								
Compliance Index	0.355	1.009	1.294	1.096	0.109	0.859***	-0.023	0.589
S.E.	(0.000)	(0.000)	(0.000)	(0.000)	(0.360)	(0.268)	(0.175)	(0.638)
N (# of teachers)	421	555	535	461	579	579	569	588
C (# of stratification cells)	19	21	20	21	20	20	20	21
Adjusted R^2	0.279	0.261	0.274	0.288	0.244	0.264	0.245	0.269
Panel C: Effect on EGRA Scores (2SLS)								
Compliance Index	0.147*	1.224**	0.633***	1.356	0.732**	0.863**	3.355	29.935
S.E.	(0.087)	(0.479)	(0.236)	(0.911)	(0.329)	(0.435)	(5.352)	(86.681)
N (# students)	988	1,255	1,235	1,161	1,191	1,191	1,221	1,238
C (# of stratification cells)	20	21	21	21	21	21	21	21
Adjusted R^2	0.188	0.195	0.190	0.130	0.155	0.148	-4.599	-23.168
Control-group values								
Mean	-0.216	0.015	0.000	0.000	0.101	0.064	2.883	1.940
SD	0.914	0.088	0.000	0.000	0.302	0.245	0.853	0.238

Notes: Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Columns 2, 3, and 4 have values mechanically set to 0 for the control group. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Columns (2)-(4) are mechanically set to 0 for control schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

C A/B Test

Table C1

A/B Test Reading Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
Treatment Effect	0.621	0.132	0.043	0.041	1.824	0.090	1.486	0.074	-0.123	-0.034	2.106	0.158	1.398	0.157	2.932	0.150	0.152	0.146
S.E.	(1.321)	(0.281)	(0.293)	(0.279)	(4.295)	(0.213)	(4.054)	(0.203)	(0.732)	(0.200)	(3.626)	(0.272)	(1.728)	(0.194)	(4.923)	(0.252)	(0.281)	(0.270)
R.I. p-value	[0.695]		[0.899]		[0.721]		[0.712]		[0.854]		[0.627]		[0.481]		[0.619]		[0.681]	
N (# students)	569		569		569		569		569		569		569		569		569	
C (# stratification cells)	38		38		38		38		38		38		38		38		38	
Adjusted R^2	0.014		0.004		0.020		0.020		0.025		0.019		-0.000		0.009		0.000	
Control-group values																		
Mean	2.084		0.785		35.295		31.095		6.594		11.828		6.849		16.314		0.526	
SD	4.704		1.053		20.203		19.954		3.652		13.338		8.913		19.564		1.038	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table C2

A/B Test Reading Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
Treatment Effect	1.908	0.418**	0.233	0.239	2.644	0.132	11.592	0.610***	2.520	0.643***	3.335	0.246	3.244	0.408***	3.415	0.164	0.159	0.164
S.E.	(0.784)	(0.172)	(0.169)	(0.174)	(3.059)	(0.153)	(2.816)	(0.148)	(0.521)	(0.133)	(2.237)	(0.165)	(1.113)	(0.140)	(3.411)	(0.164)	(0.159)	(0.164)
Treatment Leadership Effect	0.957	0.210	0.037	0.038	2.308	0.116	4.414	0.232	0.281	0.072	2.707	0.199	2.268	0.285	3.233	0.155	0.152	0.157
S.E.	(1.207)	(0.264)	(0.258)	(0.265)	(3.955)	(0.198)	(3.962)	(0.208)	(0.700)	(0.179)	(3.397)	(0.250)	(1.634)	(0.205)	(4.574)	(0.220)	(0.258)	(0.267)
N (# students)	1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298	
C (# stratification cells)	76		76		76		76		76		76		76		76		76	
Adjusted R^2	0.058		0.014		0.007		0.110		0.108		0.026		0.064		0.013		0.012	
Control-group values																		
Mean	0.649		0.621		33.533		21.441		4.754		9.438		4.061		14.289		0.434	
SD	4.566		0.974		19.984		19.015		3.920		13.580		7.960		20.817		0.967	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression with no controls. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table C3
AB Treatment Effects on Quality

		Oral Language			Phonics			Reading			Writing		
	(1) Quality Index	(2) Use New Vocab	(3) Asks Questions About Reading	(4) Practice New Language Structure	(5) Does Phonics Drills	(6) Say and Write Letters	(7) Say Correct Sounds Learners do Same	(8) Blending Sounds to Make Words	(9) Displays and Say Sight Words	(10) Read Aloud	(11) Use Writing Examples	(12) Use Class Writing Activity	(13) Gives Writing Task
Panel A: Treatment Effects on School Quality													
Treatment Assignment	0.540	0.179	0.114	0.326**	0.234***	0.058	0.046	0.305***	0.164	0.168	0.169	0.300**	0.060
S.E.	(0.455)	(0.126)	(0.106)	(0.131)	(0.069)	(0.117)	(0.114)	(0.094)	(0.135)	(0.121)	(0.106)	(0.129)	(0.147)
Panel B: Effect on EGRA Scores (OLS)													
Quality Index	0.139	0.200	0.544	0.911	0.037	0.463	-0.028	0.218	0.759	0.479	0.924	0.720	0.132
S.E.	(0.240)	(0.853)	(0.638)	(0.514)	(0.745)	(0.929)	(0.933)	(0.699)	(0.726)	(0.944)	(0.887)	(0.764)	(0.588)
N (# of teachers)	15	15	15	15	15	15	15	15	15	15	15	15	15
C (# of schools)	15	15	15	15	15	15	15	15	15	15	15	15	15
Adjusted R^2	0.449	0.426	0.448	0.540	0.421	0.440	0.421	0.424	0.468	0.436	0.499	0.471	0.423

Notes: Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Data aggregated to the teacher level. Controls picked using double-post lasso. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table C4

AB Treatment Effects on Quality (Cont'd)

		Teacher					Student				
	(1) Quality Index	(2) Appropriate Pace	(3) Good Presence & Speaks Clearly	(4) Proactively Manages Behavior	(5) Move Around Room	(6) Support Behind Learners	(7) Familiar with Routines	(8) Engaged with Workbooks	(9) Discuss After Teacher Direction	(10) Actively Involved in Activities	(11) Stay on Task
Panel A: Treatment Effects on School Quality											
Treatment Assignment	0.540	0.017	-0.073	0.336**	0.025	-0.195	0.071	-0.055	0.255*	0.080	-0.160
S.E.	(0.455)	(0.148)	(0.065)	(0.137)	(0.139)	(0.168)	(0.127)	(0.129)	(0.134)	(0.135)	(0.133)
Panel B: Effect on EGRA Scores (OLS)											
Quality Index	0.139	0.368	0.376	-0.492	0.039	-0.274	0.067	-0.533	1.833**	1.151	1.570***
S.E.	(0.240)	(0.507)	(0.533)	(0.845)	(1.225)	(1.043)	(0.525)	(0.835)	(0.573)	(0.695)	(0.356)
N (# of teachers)	15	15	15	15	15	15	15	15	15	15	15
C (# of schools)	15	15	15	15	15	15	15	15	15	15	15
Adjusted R^2	0.449	0.434	0.427	0.444	0.421	0.428	0.421	0.445	0.544	0.519	0.660

Notes: Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Data aggregated to the teacher level. Controls picked using double-post lasso. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table C5

AB Treatment Effects on Compliance

		Teacher		Student	
	(1) Compliance Index	(2) Used Workbooks (%)	(3) Used Teacher Guide (%)	(4) Prop One - One Assessments	(5) Used Report Cards (%)
Panel A: Treatment Effects on Compliance					
Treatment Assignment	0.278	-0.053	-0.135	0.259	0.036
S.E.	(0.400)	(0.111)	(0.138)	(0.344)	(0.049)
Effective F (Olea–Pflueger)	0.43				
Panel B: Effect on EGRA Scores (OLS)					
Compliance Index	0.355	0.376	0.410	0.023	1.141
S.E.	(0.400)	(0.533)	(0.720)	(0.310)	(0.680)
N (# of teachers)	13	15	15	14	16
C (# of schools)	13	15	15	14	16
Adjusted R^2	0.524	0.427	0.433	0.550	0.554
Panel C: Effect on EGRA Scores (2SLS)					
Compliance Index	0.237	3.864	1.937	0.436	-4.577
S.E.	(1.065)	(12.405)	(3.669)	(2.473)	(12.087)
N (# students)	527	562	562	532	559
C (# of schools)	30	35	35	32	35
Adjusted R^2	-0.033	-1.032	-0.220	-0.203	-0.603
Control-group values					
Mean	1.312	0.900	0.800	3.000	1.947
SD	1.490	0.308	0.410	1.138	0.229

Notes: Panels A and B are aggregated at the teacher level, while panel C is aggregated at the student level. Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Data aggregated to the teacher level. Controls picked using double-post lasso. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

D Enumerator Randomization

Table D1

Heterogeneity in Treatment Effects by Enumerator Status (SISO vs Not)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
SISO Effect	0.954*	0.217*	0.048	0.048	3.556*	0.176*	2.009	0.100	0.827**	0.209**	2.716*	0.200*	2.588*	0.332*	3.097	0.151	0.074	0.075
S.E.	(0.276)	(0.063)	(0.035)	(0.035)	(0.779)	(0.039)	(1.082)	(0.054)	(0.126)	(0.032)	(0.712)	(0.052)	(0.713)	(0.092)	(1.544)	(0.075)	(0.084)	(0.084)
Treat Effect	2.202*	0.501*	0.151	0.151	3.338	0.165	13.621***	0.678***	2.631**	0.665**	4.380	0.323	4.028**	0.517**	4.424	0.215	0.205	0.207
S.E.	(0.667)	(0.152)	(0.129)	(0.129)	(2.373)	(0.117)	(1.842)	(0.092)	(0.420)	(0.106)	(1.959)	(0.144)	(0.922)	(0.118)	(2.950)	(0.144)	(0.158)	(0.159)
SISO × Treat	0.116	0.026	0.087	0.087	-4.003	-0.198	-2.238	-0.111	0.378	0.095	-1.088	-0.080	0.316	0.041	1.038	0.051	0.238	0.239
S.E.	(0.824)	(0.187)	(0.164)	(0.164)	(4.590)	(0.227)	(5.449)	(0.271)	(0.407)	(0.103)	(2.408)	(0.178)	(1.624)	(0.209)	(2.719)	(0.132)	(0.111)	(0.111)
N (# students)	1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298	
Adjusted R^2	0.131		0.137		0.104		0.123		0.121		0.078		0.072		0.072		0.052	
Control-group values																		
Mean	0.855		0.642		33.620		22.956		4.925		9.801		4.414		14.709		0.465	
SD	4.630		0.997		20.217		20.103		3.958		13.565		7.785		20.535		0.994	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by each enumerator at a school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table D2

Heterogeneity in Treatment Effects by Enumerator Status (Teacher vs Not)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
Teacher Effect	0.190	0.043	0.144	0.149	1.651	0.081	0.574	0.029	0.128	0.032	0.280	0.022	0.183	0.023	-0.123	-0.006	-0.015	-0.014
S.E.	(0.081)	(0.018)	(0.050)	(0.051)	(0.542)	(0.027)	(0.458)	(0.023)	(0.100)	(0.025)	(0.308)	(0.024)	(0.316)	(0.040)	(0.541)	(0.027)	(0.014)	(0.014)
Treat Effect	1.975*	0.449*	0.155	0.160	3.360	0.166	10.725**	0.539**	2.555**	0.643**	4.012	0.310	3.773**	0.474**	3.201	0.160	0.204	0.200
S.E.	(0.697)	(0.158)	(0.163)	(0.169)	(2.466)	(0.122)	(2.202)	(0.111)	(0.577)	(0.145)	(1.883)	(0.145)	(0.918)	(0.115)	(2.797)	(0.140)	(0.164)	(0.161)
Teacher × Treat	0.459	0.104	0.004	0.004	-0.986	-0.049	5.057	0.254	0.231	0.058	0.475	0.037	0.571	0.072	2.605	0.131	0.055	0.054
S.E.	(0.268)	(0.061)	(0.097)	(0.100)	(1.017)	(0.050)	(2.526)	(0.127)	(0.606)	(0.153)	(0.860)	(0.066)	(0.866)	(0.109)	(1.227)	(0.062)	(0.068)	(0.067)
N (# students)	1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298		1,298	
Adjusted R^2	0.127		0.142		0.102		0.122		0.117		0.074		0.062		0.069		0.052	
Control-group values																		
Mean	0.979		0.587		33.612		23.349		5.030		10.267		4.752		15.573		0.495	
SD	4.596		0.965		20.298		19.915		3.973		12.958		7.961		19.944		1.020	

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, indicator for gender, continuous age, and their interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by each enumerator at a school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

E Survey Results

Table E1

Treatment Effects on Student Aspirations

	(1) Perceived at Bottom of Class Reading Level (%)	(2) Perceived at Bottom of Class Math Level (%)	(3) Believe will Pass HS Entry Exam (%)	(4) Believe will have Desired Job (%)	(5) Ambition Rating (1-5 Scale)	(6) Desired Job Matches Gender (%)
Treatment Effect	-0.036*	-0.040*	-0.017	0.004	-0.096	-0.012
S.E.	(0.015)	(0.020)	(0.039)	(0.005)	(0.135)	(0.028)
R.I. p-value	[0.097]	[0.061]	[0.544]	[0.422]	[0.506]	[0.729]
N (# students)	1,253	1,264	1,153	1,207	1,298	1,270
C (# stratification cells)	21	21	21	21	21	21
Adjusted R^2	0.032	0.014	0.024	0.032	0.129	0.122
Control-group values						
Mean	0.133	0.147	0.911	0.989	3.379	0.602
SD	0.340	0.354	0.285	0.103	1.144	0.490

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table E2

Treatment Effects on Teacher Practices

	(1) Enjoy Teaching English Literacy	(2) Confident Teaching English Literacy	(3) Freedom to Adjust Lessons	(4) Hours Preparing Class
Treatment Effect	0.161	0.223*	-0.077	-1.405
S.E.	(0.133)	(0.149)	(0.210)	(1.808)
R.I. p-value	[0.230]	[0.098]	[0.624]	[0.199]
N (# teachers)	70	70	67	59
C (# stratification cells)	21	21	21	21
Adjusted R^2	0.140	0.126	0.119	-0.135
Control-group values				
Mean	3.314	3.265	3.121	2.931
SD	0.471	0.448	0.600	4.303

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, fixed effects for gender, continuous age, and interactions between the two, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table E3

Treatment Effects on Student Home-Based Schooling Practices

	(1) Practice Writing at Home (%)	(2) Practice Reading at Home (%)	(3) Practice Math at Home (%)	(4) Mother Helps with Homework (%)	(5) Father Helps with Homework (%)	(6) Siblings Help with Homework (%)	(7) Believe do Better with Help (%)	(8) How often do Parents Read with Pupil
Treatment Effect	0.024	0.075**	0.020	-0.001	0.004	0.028	-0.037**	0.003
S.E.	(0.022)	(0.040)	(0.057)	(0.027)	(0.051)	(0.028)	(0.019)	(0.136)
R.I. p-value	[0.207]	[0.029]	[0.668]	[0.979]	[0.917]	[0.379]	[0.043]	[0.981]
N (# students)	1,291	1,288	1,294	1,298	1,293	1,271	1,240	1,259
C (# stratification cells)	21	21	21	21	21	21	21	21
Adjusted R^2	0.014	0.031	0.015	0.010	0.000	0.024	0.026	0.022
Control-group values								
Mean	0.903	0.774	0.687	0.693	0.576	0.764	0.910	1.476
SD	0.297	0.418	0.464	0.462	0.495	0.425	0.286	1.171

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Table E4

Treatment Effects on Student Beliefs on School Quality

	(1) Pupil Belief Index	(2) How Often Understand Teacher	(3) Everyone Knows What to Do (%)	(4) Classmates Obey Teacher (%)	(5) Learn a lot Everyday (%)	(6) Teacher Encourages Students (%)	(7) Teacher Explains Clearly (%)	(8) Pupil Enjoys School (%)
Treatment Effect	-0.045	-0.057	-0.036	-0.018	0.011	0.020	-0.011	-0.010
S.E.	(0.084)	(0.093)	(0.042)	(0.037)	(0.022)	(0.016)	(0.013)	(0.007)
R.I. p-value	[0.626]	[0.645]	[0.361]	[0.685]	[0.554]	[0.303]	[0.470]	[0.358]
N (# students)	1,230	1,277	1,264	1,281	1,279	1,285	1,287	1,296
C (# stratification cells)	21	21	21	21	21	21	21	21
Adjusted R^2	0.000	0.032	0.009	0.009	0.004	-0.005	0.005	0.011
Control-group values								
Mean	-0.000	1.681	0.812	0.740	0.928	0.892	0.950	0.980
SD	1.000	0.920	0.391	0.439	0.259	0.310	0.219	0.140

Notes: Sample is all students who were enrolled in one of the 80 study schools at the beginning of the 2024-25 academic year and were found for endline exams in June 2025. Four schools closed during the school year (two per study arm), leaving a final sample of 76 schools. Treatment effects in are estimated using a linear regression of the outcome on the treatment indicator, a complete set of age-category-by-sex interactions, and a vector of stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

F Uganda Scale up

Table F1

Observational Estimates of the Effects of the Intervention on Reading Scores in Uganda

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Overall Reading PCA Index		Listening Comprehension		Letter Names		Letter Sounds		Initial Sound Identification		Familiar Word Reading		Non-word Reading		Oral Reading Fluency		Reading Comprehension	
	Equiv. Yrs. of Schooling	SDs	Score [0-5]	SDs	CLPM	SDs	CLPM	SDs	Score [0-10]	SDs	CWPM	SDs	CWPM	SDs	CWPM	SDs	Score [0-5]	SDs
Panel A: Overall Scores																		
Treatment Effect	3.338	0.514	0.124	0.617	-3.804	-0.238	6.327	1.132	1.493	0.526	1.818	0.326	0.731	0.171	1.589	0.157	0.157	0.343
S.E.	(1.860)	(0.286)	(0.067)	(0.336)	(2.199)	(0.138)	(1.683)	(0.301)	(0.558)	(0.197)	(1.332)	(0.239)	(0.796)	(0.186)	(2.010)	(0.199)	(0.124)	(0.271)
R.I. p-value	[0.143]		[0.110]		[0.166]		[0.004]***		[0.033]**		[0.254]		[0.410]		[0.531]		[0.269]	
Control-group values																		
Mean	-0.000		0.042		23.204		4.346		1.738		2.885		1.649		5.157		0.162	
SD	6.500		0.201		15.964		5.589		2.837		5.574		4.271		10.098		0.459	
Adjusted R^2	0.154		0.148		0.051		0.177		0.134		0.124		0.067		0.124		0.063	
Panel B: Zero Score Students																		
Treatment Effect (% Change)	0.006		-0.079		-0.005		-0.225		-0.173		-0.053		-0.093		-0.156		-0.080	
S.E.	(0.006)		(0.045)		(0.028)		(0.064)		(0.074)		(0.065)		(0.062)		(0.092)		(0.070)	
R.I. p-value	[0.484]		[0.137]		[0.880]		[0.004]***		[0.050]*		[0.522]		[0.208]		[0.143]		[0.251]	
Control-group values																		
Mean	0.000		0.958		0.099		0.419		0.597		0.634		0.812		0.607		0.869	
SD	0.000		0.201		0.300		0.495		0.492		0.483		0.392		0.490		0.338	
N (# students)	371		371		371		371		371		371		371		371		371	
C (# stratification cells)	39		39		39		39		39		39		39		39		39	
Adjusted R^2	-0.009		0.106		0.003		0.100		0.067		0.104		0.056		0.191		0.067	

Notes: Sample is students enrolled in of of the 39 study schools during the 2025 academic year. Panel A shows the overall scores. Panel B shows the change in the proportion of zero scores which are defined as a binary variables that equals 1 if a student scores 0 on that EGRA component. Overall Reading PCA index is a weighted average of all the other components, where the weights correspond to the first principal component of control-group test scores. EYS stands for Equivariant Years of Schooling and is equal to the treatment effect in SDs divided by 0.22 (Evans and Yuan 2022). CLPM is correct letters per minute and CWPM is correct words per minute; both are calculated as the score on the respective subtest divided by the time taken. SDs are measured in control-group standard deviations. Treatment effects are estimated using a linear regression of the outcome on the treatment indicator, a complete set of gender-enumerator interactions. Heteroskedasticity-robust standard errors, clustered by stratification cell, in parentheses (). Randomization-inference p-values, clustered by school and using 1,000 permutations, in square brackets [ ]: *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

G Proofs

Proof of Proposition 1 The teacher chooses $I_1^A \in [0,1]$ to maximize period-2 advanced skills:

$$A_2 = A_0 + \beta \cdot I_1^A \cdot h(B_0) + \beta \cdot I_2^A \cdot h(B_1) \tag{7}$$

where $B_1 = B_0 + \alpha(1 - I_1^A)$. By backwards induction, $I_2^A = 1$ because basic skills are worthless in period 2.

Substituting the constraint on B_1:

$$A_2 = A_0 + \beta \cdot I_1^A \cdot h(B_0) + \beta \cdot h(B_0 + \alpha(1 - I_1^A)) \tag{8}$$

Taking the derivative with respect to I_1^A:

$$\frac{\partial A_2}{\partial I_1^A} = \beta \cdot h(B_0) + \beta \cdot h'(B_0 + \alpha(1 - I_1^A)) \cdot (-\alpha) \tag{9}$$

Simplifying:

$$\frac{\partial A_2}{\partial I_1^A} = \beta \left[h(B_0) - \alpha \cdot h'(B_0 + \alpha(1 - I_1^A))\right] \tag{10}$$

The second-order condition is:

$$\frac{\partial^2 A_2}{\partial (I_1^A)^2} = \beta \cdot \alpha^2 \cdot h''(B_0 + \alpha(1 - I_1^A)) < 0 \tag{11}$$

which holds since $h''(B) < 0$ by assumption. Thus, the objective function is strictly concave in I_1^A, and any stationary point is a global maximum.

Case 1: Corner solution at $I_1^A = 0$.

Since the objective is concave, $I_1^A = 0$ is optimal if and only if:

$$\left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=0} \leq 0 \tag{12}$$

Evaluating at $I_1^A = 0$:

$$\left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=0} = \beta\left[h(B_0) - \alpha \cdot h'(B_0 + \alpha)\right] \tag{13}$$

Therefore, $I_1^A = 0$ is optimal if and only if:

$$h(B_0) \leq \alpha \cdot h'(B_0 + \alpha) \tag{14}$$

Case 2: Interior solution $I_1^A \in (0, 1)$.

An interior solution exists if and only if:

$$\left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=0} > 0 \quad \text{and} \quad \left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=1} < 0 \tag{15}$$

The first condition gives:

$$h(B_0) - \alpha \cdot h'(B_0 + \alpha) > 0 \quad \Longleftrightarrow \quad h(B_0) > \alpha \cdot h'(B_0 + \alpha) \tag{16}$$

The second condition gives:

$$h(B_0) - \alpha \cdot h'(B_0) < 0 \quad \Longleftrightarrow \quad h(B_0) < \alpha \cdot h'(B_0) \tag{17}$$

Thus, an interior solution exists if and only if:

$$\alpha \cdot h'(B_0 + \alpha) < h(B_0) < \alpha \cdot h'(B_0) \tag{18}$$

The optimal interior value I_1^A is characterized by the first-order condition:

$$\frac{\partial A_2}{\partial I_1^A} = 0 \quad \Longleftrightarrow \quad h(B_0) = \alpha \cdot h'(B_0 + \alpha(1 - I_1^A)) \tag{19}$$

Monotonicity of interior solution:

For $B_0 \in [\underline{B}, \overline{B}]$, the interior solution $I_1^A(B_0)$ is implicitly defined by:

$$\Psi(B_0, I_1^A) \equiv h(B_0) - \alpha \cdot h'(B_0 + \alpha(1 - I_1^A)) = 0 \tag{20}$$

By the implicit function theorem:

$$\frac{dI_1^A}{dB_0} = -\frac{\partial\Psi/\partial B_0}{\partial\Psi/\partial I_1^A} \tag{21}$$

Computing the partial derivatives:

$$\frac{\partial\Psi}{\partial B_0} = h'(B_0) - \alpha \cdot h''(B_0 + \alpha(1 - I_1^A)) \tag{22}$$

$$\frac{\partial\Psi}{\partial I_1^A} = -\alpha \cdot h''(B_0 + \alpha(1 - I_1^A)) \cdot (-\alpha) = \alpha^2 \cdot h''(B_0 + \alpha(1 - I_1^A)) \tag{23}$$

Since $h'(B) > 0$ and $h''(B) < 0$ for all B:

- $\frac{\partial\Psi}{\partial B_0} = h'(B_0) - \alpha \cdot h''(B_0 + \alpha(1 - I_1^A)) > 0$ (positive minus negative)
- $\frac{\partial\Psi}{\partial I_1^A} = \alpha^2 \cdot h''(B_0 + \alpha(1 - I_1^A)) < 0$ (positive times negative)

Therefore:

$$\frac{dI_1^A}{dB_0} = -\frac{\partial\Psi/\partial B_0}{\partial\Psi/\partial I_1^A} = -\frac{(+)}{(-)} > 0 \tag{24}$$

This establishes that $I_1^A(B_0)$ is strictly increasing in B_0 over the interior region $[\underline{B}, \overline{B}]$.

Case 3: Corner solution at $I_1^A = 1$.

By strict concavity, $I_1^A = 1$ is optimal if and only if:

$$\left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=1} \geq 0 \tag{25}$$

Evaluating at $I_1^A = 1$:

$$\left.\frac{\partial A_2}{\partial I_1^A}\right|_{I_1^A=1} = \beta\,[h(B_0) - \alpha \cdot h'(B_0)] \tag{26}$$

Therefore, $I_1^A = 1$ is optimal if and only if:

$$h(B_0) \geq \alpha \cdot h'(B_0) \tag{27}$$

Characterization via threshold values.

Define $\underline{B}$ and $\overline{B}$ as the unique solutions to:

$$h(\underline{B}) = \alpha \cdot h'(\underline{B} + \alpha) \tag{28}$$

$$h(\overline{B}) = \alpha \cdot h'(\overline{B}) \tag{29}$$

Existence and uniqueness of $\underline{B}$:

Define $G_L(B) = h(B) - \alpha \cdot h'(B + \alpha)$. Then:

- $G_L(0) = h(0) - \alpha \cdot h'(\alpha) = 0 - \alpha \cdot h'(\alpha) < 0$ (since $h(0) = 0$ and $h'(\alpha) > 0$)
- $G_L(B) \to 1 - \alpha \cdot 0 = 1 > 0$ as $B \to \infty$ (since $h(B) \to 1$ and $h'(B) \to 0$)
- $G_L'(B) = h'(B) - \alpha \cdot h''(B + \alpha) > 0$ for all B (since $h'(B) > 0$ and $h''(B) < 0$)

By the intermediate value theorem and monotonicity of G_L, there exists a unique $\underline{B} > 0$ such that $G_L(\underline{B}) = 0$.

Existence and uniqueness of $\overline{B}$:

Define $G_U(B) = h(B) - \alpha \cdot h'(B)$. Then:

- $G_U(0) = h(0) - \alpha \cdot h'(0) = 0 - \alpha \cdot h'(0) < 0$
- $G_U(B) \to 1 - \alpha \cdot 0 = 1 > 0$ as $B \to \infty$
- $G_U'(B) = h'(B) - \alpha \cdot h''(B) > 0$ for all B

Similarly, there exists a unique $\overline{B} > 0$ such that $G_U(\overline{B}) = 0$.

Ordering of thresholds:

Note that for any $B \geq 0$:

$$h'(B + \alpha) < h'(B) \tag{30}$$

since h' is strictly decreasing (as $h'' < 0$). Therefore:

$$\alpha \cdot h'(B + \alpha) < \alpha \cdot h'(B) \tag{31}$$

This implies $G_L(B) > G_U(B)$ for all $B > 0$. Since both functions are strictly increasing and cross zero exactly once, we have $\underline{B} < \overline{B}$.

Combining the three cases with the threshold characterization:

- If $B_0 < \underline{B}$: then $G_L(B_0) < 0$, so $h(B_0) < \alpha \cdot h'(B_0 + \alpha)$, hence $I_1^A = 0$
- If $\underline{B} \leq B_0 \leq \overline{B}$: then $G_L(B_0) \geq 0$ and $G_U(B_0) \leq 0$, so $\alpha \cdot h'(B_0 + \alpha) \leq h(B_0) \leq \alpha \cdot h'(B_0)$, hence interior solution with $I_1^A(B_0)$ strictly increasing in B_0
- If $B_0 > \overline{B}$: then $G_U(B_0) > 0$, so $h(B_0) > \alpha \cdot h'(B_0)$, hence $I_1^A = 1$

This completes the proof.

Proof of Proposition 2 [Proof of Proposition 2] Define

$$g(I, B_0, S) = \beta S\Big(Ih(B_0) + h(B_0 + \alpha S(1 - I))\Big).$$

Then

$$V^{TI}(B_0, S) = \max_{I \in [0,1]} g(I, B_0, S) = g(I^*, B_0, S), \quad V^{uniform}(B_0, S, \bar{I}) = g(\bar{I}, B_0, S).$$

By the envelope theorem,

$$\frac{\partial V^{TI}}{\partial S} = \frac{\partial g}{\partial S}\Big|_{I=I^*}, \qquad \frac{\partial V^{uniform}}{\partial S} = \frac{\partial g}{\partial S}\Big|_{I=\bar{I}},$$

where

$$\frac{\partial g}{\partial S}\Big|_{I} = \beta\Big[Ih(B_0) + h(B_1) + S\alpha(1 - I)h'(B_1)\Big],$$

with $B_1 = B_0 + \alpha S(1-I)$. Define $\Delta_S := \frac{1}{\beta}\left(\frac{\partial V^{TI}}{\partial S} - \frac{\partial V^{uniform}}{\partial S}\right)$. Then

$$\begin{aligned}\Delta_S = (I^* - \bar{I})h(B_0) + \Big(h(B_1^*) - h(\bar{B}_1)\Big) \\ + S\alpha\Big((1-I^*)h'(B_1^*) - (1-\bar{I})h'(\bar{B}_1)\Big),\end{aligned}$$

where $B_1^* = B_0 + \alpha S(1-I^*)$ and $\bar{B}_1 = B_0 + \alpha S(1-\bar{I})$. At an interior optimum $I^* \in (0,1)$, the first-order condition gives $h(B_0) = \alpha S h'(B_1^*)$. Substituting:

$$\Delta_S = \Big(h(B_1^*) - h(\bar{B}_1)\Big) + S\alpha(1-\bar{I})\Big(h'(B_1^*) - h'(\bar{B}_1)\Big).$$

Suppose without loss of generality that $I^* > \bar{I}$ (the case $I^* < \bar{I}$ is symmetric). Then $B_1^* < \bar{B}_1$, so by the mean value theorem there exists $\xi \in (B_1^*, \bar{B}_1)$ such that

$$h(B_1^*) - h(\bar{B}_1) = -h'(\xi)\cdot \alpha S(\bar{I} - I^*).$$

Thus

$$\Delta_S = S\alpha\Big[h'(B_1^*) - (1-\bar{I})h'(\bar{B}_1) - (\bar{I} - I^*)h'(\xi)\Big].$$

Since $\bar{I} < I^*$, suboptimality of $\bar{I}$ requires $\frac{\partial g}{\partial I}\Big|_{\bar{I}} > 0$, which gives $h(B_0) > \alpha S h'(\bar{B}_1)$. Combined with the FOC $h(B_0) = \alpha S h'(B_1^*)$, this yields $h'(B_1^*) > h'(\bar{B}_1)$. Since h' is strictly decreasing and $B_1^* < \xi < \bar{B}_1$, we have $h'(B_1^*) > h'(\xi) > h'(\bar{B}_1)$. Therefore

$$\begin{aligned}\Delta_S &> S\alpha\Big[h'(B_1^*) - (1-\bar{I})h'(\bar{B}_1) - (\bar{I} - I^*)h'(B_1^*)\Big] \\ &= S\alpha\Big[(1-\bar{I}+I^*)h'(B_1^*) - (1-\bar{I})h'(\bar{B}_1)\Big] \\ &> S\alpha(1-\bar{I})\Big[h'(B_1^*) - h'(\bar{B}_1)\Big] > 0.\end{aligned}$$

For corner solutions at $I^* = 0$ or $I^* = 1$, replace the FOC with the appropriate inequality. The same argument shows $\Delta_S > 0$ whenever $\bar{I} \neq I^*$. Therefore,

$$\frac{\partial V^{TI}}{\partial S} > \frac{\partial V^{uniform}}{\partial S},$$

establishing that targeted instruction and quality improvements are complements.

H TFLI Materials Examples

Figure H1

Teaching Guide Example Page 1

Week 7 | Day 2

Things you'll cover today...

New Vocab	Review Sound	Sight Words	Writing Skill
building trading teaching	e	she of	Full Stops

1 Oral Language — 10 min

For today's oral language we are going to learn and practise using vocabulary relating to our theme, 'Activities in Our Community'.

Warm up with a Song — 2 min

I will sing each line of the song, and you're going to sing each line back to me.

Cast, Cast, Cast

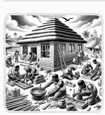
To the tune of 'Row, Row, Row Your Boat'

Cast, cast, cast your net,
Gently on the sea,
Happily, happily, happily, happily,
Catch some fish for me.

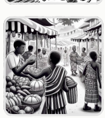
1.1 Teach New Vocabulary — 4 min

Turn to page 99 of your workbook. Let's say today's new words and learn what they mean.

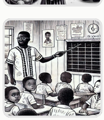
New Vocabulary



building Verb
The action or trade of constructing something, like a house.



trading Verb
The action or activity of buying and selling goods and services.



teaching Verb
The action of helping someone learn new things.

I want you to think of a sentence using one of our new words. For example, "The builder was building a new house."

I want you to talk in pairs and each share two sentences using the new words. I will be moving round the room to listen...

Circulate to check learners are using the new words correctly in sentences.

I would like <u>two</u> pairs to share their sentences with the whole class...

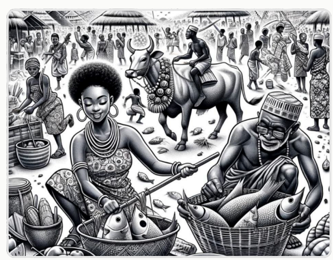
1.2 Interactive Read Aloud — 4 min

I am going to read out a story and I want you all to look at the big picture on page 99 in your workbook. The title is, 'The Big Feast'.

1. What do you think is taking place in the picture?

> Read the text aloud with feeling and expression

The Big Feast



In a busy village, everyone was getting ready for the big cultural festival. Kwame the fisherman, Kojo the cattle keeper, and Esi the farmer decided to work together to provide food for the community. Kwame went to the sea to catch fish, Kojo cared for his cattle, and Esi harvested fruits and vegetables from her farm. On the festival day, the village was full of people and music. The three shared their food, and there was singing, dancing, and eating. The festival brought the village together, showing that when people work together, they can make something special.

I have some questions to ask about the story.

2. What did Esi, Kojo and Kwame provide for the community?
3. Why were the people dancing and singing?

BS1 | Term 1   130   Week 7 | Day 2

Figure H2

Teaching Guide Example Page 2

2 Phonics — 20 min

Next, we are going to practise our phonics and review our recently learnt sound, /e/.

2.1 Which Word Starts With /e/ — 2 min

I am going to say two words.

Shout the word that starts with an /e/ sound.

Say each word pair

red	engine	empty	table
end	hand	bus	echo

2.2 Odd One Out — 2 min

I'm going to say 3 words. I want you to tell me which word does not rhyme.

Listen carefully and tell me which one is the odd one out...

Say each set of 3 words...

1.	net	mat	sat
2.	run	sun	sand
3.	log	book	dog

2.3 Oral Segmenting Drill — 2 min

Next, we are going to say the word and then break it up into the sounds.

As I say a word, I want you to repeat the word, then say the sounds that are in it, like this, "top, /t/ /o/ /p/, top".

Say each word once

set met den test

Do not segment the sounds for the learners.

2.4 Auditory Drill — 3 min

Open your workbooks to page 100. When I say the sound, you write the letter.

Say each sound once

/d/ /e/ /t/ /a/

2.5 Review the sound /e/ — 8 min

Yesterday we learnt the sound /e/.

My turn, /e/. Your turn...

Say each word once

e

The letter 'e' makes the sound /e/.

My turn, /e/, your turn... Again, /e/.

Stand up and put your hands together in the air. Let's air-write 'e'.

As we do it, say it with me, "across, up, around."

OK, open your workbook to page 100. It is your turn to practise writing 'e'.

Learners practise writing it

Check learners are writing correctly and support learners with their tripod grip.

Turn to page 100 in your workbooks.

As a whole class, let's read the words and a sentence that use the sound /e/.

BS1 | Term 1 131 Week 7 | Day 2

Figure H3

Teaching Guide Example Page 3

Practise reading with it

bet	ten	sent	bed

He gets ten pens.

Now, read the words and sentence again quietly in your pairs.

Move around the room listening to pairs reading.

3 Reading — 15 min

For our reading today, we are going to practise blending and review our sight words.

3.1 Blending Drill — 8 min

Turn to page 100 in your workbook. Put your finger on the first word chain. Let's read together...

Word Chains

den > hen > ten > pen

fin > fig > fog > frog

Ted > bed > bend > send

Next, I would like you to read them through again quietly in your pairs...

3.2 Review Sight Words — 7 min

Let's practise reading the new sight words we learnt yesterday and remind ourselves of how they sound, and what they mean...

Recent Sight Words

s h e   o f

Turn to page 100 in your workbook.

We are going to read through some sight words we've learnt recently together.

Sight Word Practice

so	of	by	on
for	can	not	she

Now I want you to read them again, quietly with your partner.

4 Writing — 15 min

For our writing session today we are going to learn about full stops.

4.1 Teach Full Stops — 5 min

We have learnt that sentences start with a capital letter.

Today we will learn what sentences end with.

Show & explain a full stop

I pat the mat.

I have circled the full stop. It is at the end of my sentence.

Full stops are very important. They tell us when a sentence has ended.

This helps to separate our sentences so that we can understand them.

4.2 Shared Writing Practice — 5 min

> Write out the sentences without full stops.

Example (without full stops)

Sam is a man Pat is an ant

Let's read these sentences together.

It is hard to understand the sentences because the full stop is missing.

BS1 | Term 1 — 132 — Week 7 | Day 2

Figure H4

Teaching Guide Example Page 4

I know we start a sentence with a capital letter so my full stop must go before it.

Example (with full stops)

Sam is a man. Pat is an ant.

Let's read the sentences again and add the full stops together. "Sam is a man, Pat is an ant"

Where do the full stops need to go?

Yes, that's right. One after man. And another at the end, after ant. Why is that?

4.3 Independent Writing

5 min

Now, turn to page 101 in your workbook.

I want you to read the text and circle all the full stops you can see.

Check all learners are circling the full stops.

Now, look at the next text in your workbook.

What do you think is missing?

That's right, full stops.

I want you to spend 3 minutes adding in all the missing full stops.

Use the capital letters to help you.

Sam and the Dog

Sam is on a mat

A dog is on a log Dad pats the dog

The dog did a spin Sam pats dog

Dad and Sam sit on the mat

Check all learners are circling the full stops.

What do you notice comes after a full stop?

Lesson Recap

2 min

> Ask the class these questions to recap the lesson.

Let's review what we've learnt today...

Oral	Have you seen someone building a house? What did you see?
Phonics	Is there an /e/ sound in the word 'net'?
Reading	We read a story. What happened in the village?
Writing	How do you know a sentence is finished?

✓ End of lesson

BS1 | Term 1 133 Week 7 | Day 2

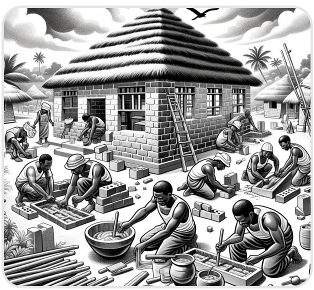
Figure H5
Workbook Example Page 1

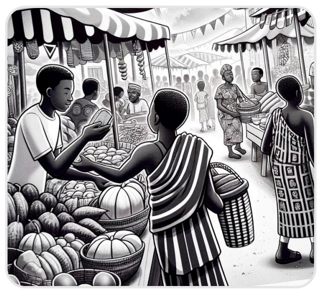
Week 7 | Day 2 | Our Community

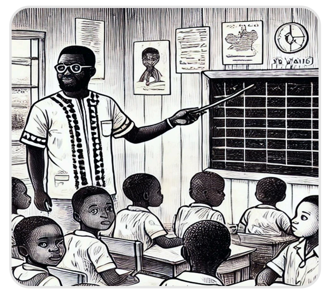
Activities in our Community

Oral Language

1. Talk about the new vocabulary:


building


trading


teaching

2. Read Aloud - Talk about the picture:

Figure H6
Workbook Example Page 2

Phonics

1. Write the letter sound you hear:

2. Write e:

3. Read words with e:

bet	ten	sent	bed

He gets ten pens.

Reading

1. Read the word chains:

den > hen > ten > pen

fin > fig > fog > frog

Ted > bed > bend > send

2. Read recent Sight Words:

so	of	by	on
for	can	not	she

BS1 | Term 1 100 Week 7 | Day 2

Figure H7
Workbook Example Page 3

Writing

1. Circle the full stops:

Sam is a man . Pat is an ant .

"It is my pan . It is a tin pan ."

Pat is in the pan . "Spin me," said Pat .

Sam spins Pat in the pan .

2. Add the missing full stops:

Sam is on a mat

A dog is on a log Dad pats the dog

The dog did a spin Sam pats dog

Dad and Sam sit on the mat