NBER WORKING PAPER SERIES

EXPERIMENTAL EVIDENCE ON ALTERNATIVE POLICIES TO INCREASE LEARNING AT SCALE

Annie Duflo
Jessica Kiessel
Adrienne Lucas

Working Paper 27298
http://www.nber.org/papers/w27298

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2020, Revised March 2021

We gratefully acknowledge generous funding for the evaluation from the International Growth Centre, the Hewlett Foundation, and the Children's Investment Fund Foundation. Many thanks to the Amma Aboagye, Albert Akoubila, and Maame Araba Nketsiah for supporting and championing the implementation of program and to Ama Anaman, Raphael Bandim, Suvojit Chattopadhyay, Callie Lowenstein, Sam N'tsua, and Pace Phillips for outstanding research implementation and project management. We would also like to thank Wendy Abt for her instrumental role in getting this project started and Caitlin Tulloch and Shahana Hirji for their leadership and support with the cost analysis. For research assistance, we thank Joyce Jumpah, Ryan Knight, Harrison Diamond Pollock, and Matthew White. We also acknowledge our partners at the Ministry of Education, Ghana Education Services, and the Ministry of Youth Sports and Culture without whom this project would not have been possible. We thank David Evans and Fei Yuan for providing statistics on existing impact evaluations and James Berry for providing combined math and literacy estimates for the interventions in Banerjee et al. (2017). For useful comments and suggestions, we thank Noam Angrist, Sabrin Beg, Jim Berry, Janet Currie, David Evans, Anne Fitzpatrick, John Floretta, Alejandro Ganimian, Sarah Kabay, Heidi McAnnally-Linz, Daniel Rodriguez-Segura, Jeremy Tobacman, and seminar participants at Swarthmore College, the University of Delaware, the Northeast Universities Development Consortium Conference, and the Teaching at the Right Level Conference. This RCT was registered in the American Economic Association Registry for randomized control trials as AEARCTR-0005912. The Innovations for Poverty Action IRB approved this study. This paper was previously circulated under the titles “Every Child Counts: Adapting and Evaluating Targeted Instruction Approaches into a New Context through a Nationwide Randomized Experiment in Ghana” and “External Validity: Four Models of Improving Student Achievement.” The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2020 by Annie Duflo, Jessica Kiessel, and Adrienne Lucas. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Experimental Evidence on Alternative Policies to Increase Learning at Scale
Annie Duflo, Jessica Kiessel, and Adrienne Lucas
NBER Working Paper No. 27298
June 2020, Revised March 2021
JEL No. I21,I25,I28,J24,O15

ABSTRACT

We partnered with the Ghanaian government to evaluate four methods of increasing achievement in schools with low average but heterogeneous student achievement. All methods focused teaching at the learning level of the child—a remedial pull-out program with a teaching assistant, a remedial after school program with an assistant, an assistant teaching half the students, or teachers focusing on homogeneous groups of learners. Despite imperfect implementation, student learning increased across all four, more so for female students, and gains persisted after the program ended. Fidelity of implementation decreased over time for the assistants but increased for the teachers.

Annie Duflo
Innovations for Poverty Action
aduflo@poverty-action.org

Jessica Kiessel
Omidyar Network
Redwood City, CA
jrkiessel@gmail.com

Adrienne Lucas
Lerner College of Business and Economics
University of Delaware
419 Purnell Hall
Newark, DE 19716
and NBER
alucas@udel.edu

1 Introduction

Many developing countries have eliminated the fee-based barriers to primary school enrollment, resulting in large increases in the number of children in school (Lucas and Mbiti 2012). Unfortunately, education systems originally designed for a smaller cadre of teachers to teach a more homogeneous group of students are failing to educate students in this larger, more heterogeneous environment comprised of many first generation learners. Effective solutions have been proposed through randomized controlled trials, yet they often have their own limitations and whether they can increase learning when integrated into existing systems at scale is unknown. This paper tests, in existing systems, four alternatives to support teachers' transition to the new status quo, a frontier challenge for developing countries. In a single 500 school randomized controlled trial that reached over 80,000 students, we test four models that draw on some of the most effective content delivery interventions in the last 20 years in developing countries–assistant teachers, smaller class sizes, additional instructional time, tracking, and remedial and differentiated instruction–and are the first to show their potential, relative effectiveness, or effectiveness over time when fully designed and implemented by existing government systems. Results from this study have already influenced the implementation of programs to improve education in 11 countries in Africa, with the potential to reach over 70 million children per year at scale.

Specifically we evaluated the Teacher Community Assistant Initiative (TCAI), a Ghana Ministry of Education program to increase student learning through existing schooling and youth employment systems. In each intervention, existing education sector employees designed teaching and learning materials, trained educators in student-centered, active pedagogy, and provided the educators accompanying teaching and learning materials. Three of the interventions used an existing youth employment scheme to hire teaching assistants to work with 1) remedial learners on a pull-out basis during the school day, 2) remedial learners outside of the school day, or 3) half of the classroom each day on grade-level content. The fourth intervention trained teachers to divide students by learning level for part of the day

and focus instruction on students' learning levels, i.e. differentiated instruction. We evaluated the effectiveness of each intervention by randomizing 500 schools into one of the four treatment arms or a control group and conducting 9 rounds of data collection over three school years.

All four interventions increased student achievement, showing that existing systems have the capacity to deliver more learning. After one year of exposure, the two assistant-led remedial arms increased student test scores by 0.10 standard deviations (SD). After two years of exposure, the assistant-led remedial after school program increased student test scores 0.15 SD, the assistant-led remedial during school program increased scores 0.14 SD, and both the assistant random split and the teacher-led differentiated instruction version increased scores 0.08 SD. These two year increases were equivalent to 18 to 34 percent of a year of schooling in this context. For the remedial and differentiated arms, the effects were larger for the test questions that focused on foundational content. Because the assistant arms were more expensive, but also more effective, the cost-effectiveness of the assistant-led remedial arms and the teacher-led differentiated instruction arm were approximately equivalent. While the interventions were not designed to confer differential effect sizes by gender, the test scores of female students increased by at least 0.10 SD more than for male students in the three interventions that had a remedial or differentiated component. We test the persistence of the effects on students who were treated for about one year and tested a year after the program ended. The assistant-led remedial after school and assistant split arms show persistent test score gains of about 0.10 SD while the other arms are no longer statistically significant.

Because the interventions share common elements, we use a conceptual framework to show three important mechanisms: 1) a smaller class size and an extra instructional hour are almost perfect substitutes in the presence of remedial instruction, 2) a smaller class size is more effective when paired with remedial instruction, and 3) teacher-led differentiated instruction, which included remedial elements and a more homogeneous learning environment,

was an imperfect substitute for purely remedial instruction by assistants.

Beyond learning, the program did not affect students' likelihood of being present, dropping out, or repeating a grade level, common concerns with tracking programs, or the likelihood that teachers were present at school. Teachers in the teacher-led arm were 11 percentage points more likely to be engaged with students, and conditional on being present, 5 times as likely as those in the control group to be using teaching and learning materials.

Implementation offers both optimism and caution from an entirely government run program: over 85 percent of targeted educators received training, yet fidelity of implementation was low due to material and payment delays, and absenteeism was high because the programs operated within existing incentive and supervision structures. When treatment educators should have been implementing the intervention, teachers in the teacher-led arm only divided students by learning level 6 percent of the time and assistants across all three arms worked with their smaller groups only 34 percent of the time. Adherence in the assistant-led arms fell over time from 47 percent in year 1 to 22 percent in year 3.

In addition to already influencing policy, our findings make three related contributions to the economics literature. First, our four alternatives to support teachers' transition to the new status quo incorporate some of the most promising separate findings across other interventions, providing evidence that studies with only a single intervention per context cannot.[1] The assistant-led during school remedial model built off the NGO-supported assistants in Banerjee et al. (2007) that increased learning in Mumbai and Vadodara, India and student tracking in Western Province, Kenya in Duflo et al. (2011). The assistant-after school model built from Banerjee et al. (2010), which increased letter recognition in Jaunpur district, India. In contrast to teacher-led differentiated instruction in Bihar and Uttarakhand, India, which did not increase learning, our version increased student test scores even without needing the extra supervisory layer and instructional hour inplemented in Haryana, India

[1]One potentially promising class of interventions we do not address are those using technology (see Beg et al. 2020 for a summary of the literature). Requirements of security, electricity, and internet connectivity rendered such interventions impractical in this context. Most education RCTs in lower income countries only contain one treatment arm (Evans and Yuan 2019).

(Banerjee et al. 2017). Our assistant-split intervention also contributes to the literature on contract teachers–teachers hired on a fixed term contract at a lower salary than government teachers–an alternative way to improve student achievement that bypasses existing teachers instead of supporting them with aides or training (Muralidharan and Sundararaman 2013; Duflo et al. 2015; Bold et al. 2018). Across education meta analyses, one of the most consistently recommended interventions to increase student learning has been teaching at the learning instead of grade level of the student through differentiated or remedial instruction, and we are the first to successfully implement it without substantial NGO support (Banerjee et al. 2017; Evans and Mendez Acosta 2021). All four of our models offer alternative ways to implement instruction more focused on individual learners–during or after school focusing on remedial learners, by dividing the class in half, or having existing teachers specifically focus on a more homogeneous group of learners.

Second, by comparing the effects and cost effectiveness of the four alternatives together in a new context, we further contribute to the understanding of the external validity of these methods and which is the most effective and cost-effective way to increase learning. The effect sizes of the assistant variations were smaller than the NGO-led implementations in India, confirming the findings in Vivalt (2020) that programs implemented by governments have smaller effect sizes than those implemented by an NGO or researcher. Yet, we also find an important caveat–the teacher-led model increased learning more than previous, similar models with limited NGO support. We show that across institutional contexts, providing students more tailored instruction increases learning while acknowledging that the agency in charge of implementation might matter as much as the program design.

Third, we show that existing government structures have the capacity to increase learning with sufficient political will and that sustained political will is necessary for continued success, mimicking likely effects at scale. Previous programs that embedded NGO-designed programs in existing, and hesitant, government structures did not increase student learning (Banerjee et al. 2017; Bold et al. 2018). In this study, the government involvement started at the

outset in the design of the teaching, learning, and training materials and continued through training and implementation, creating a truly government owned and operated program and demonstrating the potential potency of the program if implemented elsewhere entirely within a government system. Members of sub-entities within the Ministry of Education who understood the tension between completing the official curriculum and reaching all students created the teaching, learning, and training materials and supported its implementation. Yet, we also show that continuing political will beyond program inception is also crucial–the assistants' falling adherence to the program over time echoed the erratic payment reliability and the increasing adherence by the teachers reflected modifications made in response to teacher request.

2 Background

2.1 The Ghanaian Educational System

Primary school in Ghana is grades 1 through 6, starts at age 6, and is free of tuition fees in government schools. Our study focuses on students in government schools in grades 1-3, i.e. lower primary. The school year starts in September and consists of approximately three 13 week terms: mid-September through mid-December, January through mid-April, and May though the end of July.

In lower primary school, teachers are grade-level classroom teachers, teaching all subjects to a classroom of a specific grade-level of students. Teachers' salaries are paid centrally, and Ghana Education Service (GES) assigns teachers to schools.

Even though the number of children in schools and the heterogeneity of their family backgrounds and pre-school preparations have increased substantially since the start of free primary education in Ghana in 2005, the official curriculum to which teachers must adhere is largely unchanged from a time in which only wealthier, more highly educated parents could afford to send their children to school. Large, heterogeneous classrooms and teacher-focused

pedagogy leave many students behind–only about a quarter of primary school students reach proficiency levels in English and math (Ministry of Education 2014). In our baseline data, only 6 percent of grade 3 students could read a grade 3 text and 18 percent could not identify letters of the alphabet. As with many other lower income countries with high stakes certification exams between schooling levels, teachers are expected to adhere to a national curriculum even if students are well behind grade level. This pressure often causes them to focus on the highest achieving students, those at grade level or above (Gilligan et al. forthcoming).

In lower primary grades the language of instruction should be the school's assigned National Literacy Acceleration Program (NALAP) language. This policy started in the school year immediately preceding our study with implementation lingering into our study years (Hartwell 2010). This NALAP language was to became the primary language of instruction in grades 1-3, with instruction shifting fully to English in grade 4. We developed appropriate learning and testing tools for the 9 NALAP languages of our study schools. Because of the NALAP delays our analysis focuses on math and English skills, providing estimates for the NALAP language separately.

2.2 National Youth Employment Program

The National Youth Employment Program (NYEP) paid the intervention's assistants, known as Teacher Community Assistants (TCAs). The NYEP was an existing program under the Ministry of Youth and Sports that offered unemployed youth (18 to 35 years old), mostly secondary school graduates, two year public service positions and a small ($80-100) monthly stipend. NYEP youth were already used by the Ghana Education Service on a limited basis to fill vacant teacher positions, often in remote areas.

3 Intervention and Conceptual Framework

3.1 Intervention

The project was a partnership between GES, the Ghana National Association of Teachers, and the NYEP. In preparation for the implementation, Ghanaian education officials visited India in August, 2010 to learn about the previous successes and challenges of the Teaching at the Right Level (TaRL) approach that was implemented by Pratham, a large Indian education NGO. The government team designed the teaching, learning, and training materials with inspiration from the TaRL approach.

This study tested four models of improving student learning relative to each other and a control group—assistant-led remedial during school, assistant-led remedial after school, assistant-split, and teacher-led differentiated instruction. Treatment was assigned at the school level with 100 schools receiving each treatment. Figure 1 summarizes the components of each intervention. The interventions were not strictly nested but did contain common elements across multiple interventions.

[Figure 1 about here]

Schools in all three assistant-based treatments received the same hiring instructions: School Management Committees (SMCs) and Parent Teacher Associations (PTAs) were to identify potential assistants, secondary school graduates aged 18 to 35 living in the school community, who would be interviewed by a panel of local, GES, and NYEP representatives.

Existing government trainers trained assistants and treatment teachers in all arms on child-focused, active learning and received a bank of engaging, child focused potential activities. Assistants and teachers in the remedial and differentiated instruction arms received additional training on teaching at the learning levels of students and specific materials for each learning level, outlines of topics to cover but not scripted lessons. All assistants and teachers were responsible for their own lesson plans with the provided materials as a guide. The programs were implemented with minimal support from four Regional Coordinators

who were each responsible for 100 regionally proximate schools and reported to the Director of Basic Education.

The assistants and teachers in the remedial and differentiated arms were to test students at the start of each term to determine their learning levels. Assistants then worked with each learning level across the three lower primary grades for 4 hours per week. Initially, the differentiated instruction teachers divided students by learning level within their classrooms for one hour each day, but teachers reported that this was too hard to manage. Starting with the second term of year 2, teachers divided their students across grades by learning level for one hour each day with one teacher focusing on each level. Assistants in the assistant split worked with a random half of the students for 4 hours per week during school hours on grade level material.

The interventions occurred during three academic years. Initial trainings occurred in May (Term 3) of the 2010-2011 academic year with treatment lessons starting immediately despite material delays that lasted into the second academic year. Additional training sessions occurred throughout the next two academic years, with the study ending at the end of the 2012-2013 academic year.

The labels above the line in Figure 2 display the academic year and intervention timeline. The labels below the line are the nine data collection points.

[Figure 2 about here]

Our primary cohort of interest was subject to the intervention (or in the control group) starting with the third term of grade 1. They continued with these interventions through the end of grade 3. We further provide effects for the cohort that was subject to the intervention starting in the third term grade 2, stopped being directly subject to the interventions at the end of grade 3, and were tested again at the end of grade 4, one full year after leaving the program.

3.2 Conceptual Framework

Even though the interventions were not strictly nested, the commonalities and differences between them and their relative effect sizes are informative about mechanisms to improve student outcomes. The overall effect of each intervention relative to the control group compares the total size of the particular bundle relative to the status quo. Other comparisons provide additional insight, effectively the partial derivative from marginal changes to an intervention designed to increase student learning.

Comparing the two assistant led-remedial interventions (T1 vs T2 in Figure 1) shows the relative merits of using smaller class sizes versus an extra instructional hour to deliver remedial material. The comparison of the two during school assistant interventions (T1 vs T3) shows the marginal effect of remedial versus grade level instruction. The relative magnitudes of the assistant-led remedial during school and the teacher-led differentiated instruction interventions (T1 vs T4) shows whether a classroom teacher can replicate the benefits of a smaller class size by focusing on a more homogeneous group of learners. When comparing the assistant-led remedial after school to the assistant split (T2 vs T3), the difference is the relative benefit of remedial instruction plus an extra instructional hour relative to a smaller class size. The assistant-led remedial after school relative to the teacher-led differentiated (T2 vs T4) shows the relative merits of an extra instructional hour focused only on remedial students versus more homogeneous instruction during the normal school day. The final comparison of the assistant-split relative to the teacher-led differentiated (T3 vs T4) shows the relative effect of a smaller class size versus a more homogeneous learning environment.

4 Empirical Strategy

From our 5-armed randomization design, comparing outcomes between students in the treatment and control schools is straightforward. Formally, we estimate an intent to treat speci-

fication

$$y_{is} = \alpha + \sum_{T=1}^{4} \beta_T \, treatment_{Ts} + X_{is}' \Gamma + \varepsilon_{is} \tag{1}$$

where y_{is} is outcome y for student i in school s, $treatment_{Ts}$ is an indicator variable equal to one if school s was a treatment T school with a separate indicator for each of the four treatments (the control group is the omitted category), X_{is} are a vector of individual level controls, and ε_{is} is a cluster-robust error term assumed to be uncorrelated between schools but allowed to be correlated within a school. When the outcome of interest is a student's test score, we implement a lagged dependent variable model and include the test score from the baseline as a control in the X_{is} vector. We always include dummy variables for strata (region by above/below median pupil teacher ratio by above/below median baseline test score) and gender in X_{is}.

We test the impact of the treatment on the students' test scores, attendance, likelihood of dropping out, and likelihood of being demoted or held back a grade and on teachers' and assistants' attendance, time on task, and material usage.

5 Sample Selection and Data

The 500 school sample is from Ghana's Education Management Information Systems (EMIS) school list. The schools were selected by randomly selecting 42 out of 170 total districts with at least two from each of the ten regions, allocating each district to either receive 11 or 12 sampled schools, and then selecting those schools, attempting to get an equal number of urban and rural schools. From this sample, schools were randomly allocated into the four treatment arms and control group, stratified by region, above/below median average baseline student test score, and above/below median pupil teacher ratio. At the school level, a sample of 25 pupils from each of grades 1 and 2 was randomly selected from students present the day of enumeration, with approximately equal gender balance. We attempted to track these students for two years.

To evaluate the effect of the four interventions, we collected nine rounds of data between October 2010 and July 2013–a baseline, six spot-checks, and two achievement follow-ups. In the baseline and achievement follow-ups we administered surveys to head teachers, i.e. principals, teachers, and students and tested students using bespoke exams in all 500 schools. In the spot-checks we visited a sub-sample of schools, observing classrooms and collecting attendance data on students, teachers, head teachers, and assistants. Appendix Section A.1 contains additional details on data collection and test design.

Data from our five treatment arms are balanced based on student, teacher, school, and assistant characteristics (see Appendix Section A.2). On average each grade had 37 students and one teacher.

Baseline achievement levels were low and heterogeneous within schools. About 30% of grade 2 students could not correctly name a presented uppercase English letter, only 3% could read a three letter word, and one-third could not perform simple 1-digit addition. At baseline, the average standard deviation within a school and grade was equal to almost 90% of the average score difference between grades 1 and 2.

6 Results

6.1 Student Outcomes

Achievement

Table 1 contains the effects of the four treatments on the combined math and English student test scores, using Equation 1. The first four columns are for students who were grade 1 at baseline and should have been grade 2 at follow-up 1 and grade 3 at follow-up 2. We attempted to interview and assess all baseline students regardless of their grade-level at follow-up. After only about two terms of treatment, relative to the control group, students' test scores were about 10 percent of a standard deviation larger for the two assistant-led remedial interventions (column 1). The other two interventions had smaller, statistically

insignificant test score gains. We fail to reject statistical equality between any pair of interventions. In the grade 3 follow-up, test scores increased across all of the interventions by a statistically significant 0.08 SD (teacher-led differentiated instruction and assistant split), 0.14 SD (assistant-led remedial during school), and 0.15 SD (assistant-led remedial after school) (column 2). As with the grade 2 follow-up, we fail to reject equality between any pair of coefficients. Even though students were exposed for twice as long to the intervention at the grade 3 follow-up, the test scores gains are only about twice as large for the assistant-split–they are about 50 percent larger for the remedial and differentiated versions. To provide context for these test score gains, we compare them to the test scores across grades in control schools. As of follow-up 2, grade 3 students in control schools scored 0.41 SD higher than grade 2 students. Therefore, the two year test score gains from the treatments were equivalent to an additional 18 percent (teacher-led) to 34 percent (assistant led after school) of a grade level.

[Table 1 about here]

As two of the interventions were focused on remedial skills, in columns 3 and 4 we restrict the exam questions to foundational literacy and numeracy questions, those most similar to the Annual Status of Education Report (ASER) exam conducted in South Asia.[2] The interventions with a remedial or differentiated focus have slightly larger effects on these questions than the full exam while the assistant split has smaller effect sizes (columns 3 and 4). The gains between the two follow-up rounds are smaller, ranging from 14 to 35 percent.[3]

Across both types of questions, all four interventions increased test scores more when measured in grade 3 than in grade 2 but the rate of increase of test score decreased between the two data collection rounds. As we discuss more fully in Section 6.2, fidelity of implementation also feel between these two data collection rounds.

[2]The ASER uses four types of questions to assess a student's reading level: reading letters, words, sentences, and paragraphs. Students are not asked comprehension questions. For math, students are asked to identify one digit numbers, identify two digit numbers, perform two digit subtraction with borrowing, and division of a three-digit number by a one-digit number.

[3]Appendix Table A4 contains subject-specific test scores for both the full test (Panel A) and the foundational content only (Panel B).

In the final two columns of Table 1, we test for the persistent effects of the intervention on the cohort of students who were treated during grades 2 and 3 and tested in grade 4, one year after they aged out of the program. The knowledge on grade 1-3 material largely persisted with test scores gains of over 0.10 SD for the three assistant versions but these gains are more muted when considered along with grade 4 level material–we only found statistically significant increases for the assistant-led remedial (0.08 SD) and the assistant-split (0.12 SD) (columns 5 and 6).[4] The increase in students' specific knowledge from the intervention persisted but it did not necessarily allow students to absorb new content more readily in two of the four interventions. The focus in the assistant split on grade-level content could have prepared students better for grade 4 content than the remedial interventions. Despite initial effects that are smaller than similar versions in Banerjee et al. (2007), the persistent effects on foundational material are similar to the foundational gains for students one year removed from their balsakhi program.

Selection Into the Test

Even though attrition was about 25 percent, we found no differential selection by treatment status or the interaction of treatment status and baseline achievement (Appendix Table A6). Nevertheless, we provide Lee (2009) bounds in Appendix Table A7, finding similar results as those in Table 1.

Additional Outcomes

We use data from our unannounced spot-checks to test for the effect of the intervention on three additional outcomes, sometimes considered non-cognitive outcomes: absenteeism, no longer attending school, and grade repetition (Jackson 2018). At these spot checks about 64% of baseline students were present in control schools, 22% were no longer attending that school (the sum of dropping out and transferring), and 18% were in a grade level lower than the expected grade. None of the interventions changed these likelihoods. The full point values appear in Appendix Table A8.

[4]Appendix Table A5 contains the effects on the subject-specific tests.

6.2 Program Implementation and Time on Task

In Table 2, we present results for head teacher, teacher, and educator (either assistant or teacher depending on the intervention) time on task based on data collected during the six unannounced spot-check rounds. In both the assistant-led remedial during school and teacher-led models, head teachers were more likely to be present than in the control schools (14 and 9 percentage points, column 1). None of the interventions changed the likelihood that teachers were present at school (column 2), but teachers in the teacher-led differentiated instruction intervention were 5 percentage points more likely to be in their classroom (column 3) and 11 percentage points more likely to be engaged with students, a 31 percent increase over the control group mean of 36 percent (column 4). Conditional on a teacher being present in the classroom, only about 5 percent of control group teachers were using teaching and learning materials and teachers in the assistant-focused interventions were no different (column 5). Therefore, the emphasis on using materials did not transmit between the assistants and the classroom teachers in the schools with assistants. Teachers in the teacher-led model were 21 percentage points more likely to be using materials, a 400 percent increase over the control group.

[Table 2 about here]

In column 6 we estimate fidelity of implementation as the likelihood that group meetings were happening as they should at the school level, i.e. assistants or teachers teaching to the appropriate group of students. In all interventions the group teaching was happening statistically significantly more than in the control schools where we never observed teachers teaching to a subset of their classes. Further, we reject that the interventions were equally likely to happen. The after school lessons happened in about 40% of schools, the two in-school assistant lessons in about 30% of schools, and the teacher-led model in about 6% of schools. All schools that were supposed to hire assistants did so.

These results suggest that in the teacher-led intervention, teachers increased their use of materials and their likelihood of being engaged with students but were largely not cor-

rectly differentiating their instruction by working with more homogeneous groups of students. Therefore, students received more effective teaching, just potentially not in the full manner intended by the design of the program.

The evolution of the fidelity of implementation shows the importance of continued political will and support for the assistants and teachers beyond the initial interest. In the first year of the program, the two assistant remedial interventions occurred about 50 percent of the time, the assistant split about 40 percent of the time, and the teacher-led model a statistically insignificant 4 percent of the time. All educators received refresher trainings throughout the implementation. Assistants faced payment delays, and decreased their fidelity of implementation, while in contrast instructions around the teacher-led implementation improved in response to teacher feedback, and the fidelity of implementation increased for teachers (Table 3). By academic year 3, assistants in the assistant-led remedial after school version were only working with their groups 34 percent of the time, assistant-led remedial during school only 20 percent of the time, assistant-split 10 percent of the time, and the teacher-led differentiated instruction was at a high point of 6 percent. The fall in the fidelity of implementation was not due to assistants being absorbed as classroom teachers–they were observed covering for the classroom teacher about 6 percent of the time across all years–instead their attendance steadily fell over the course of the intervention from 69 percent in the first year to 42 percent in year 3. Therefore, schools did not fully absorb assistants as regular teachers nor did the assistants completely adhere to the intervention even when they were present.

[Table 3 about here]

6.3 Heterogeneity

The analysis thus far focused on the test scores of all students as even the remedial interventions could have helped all students by creating more homogeneous learning times during the pull-out sessions or bringing remedial learners closer to grade level. The heterogeneity

analysis in Table A9 confirms this uniformity of effect–the interaction between baseline student test score and treatment is statistically insignificant in all cases as is the interaction of treatment status times whether the student was in the top two thirds of the baseline school by grade test score distribution, an approximation of not receiving remedial instruction.

The intervention was not designed to favor one gender, yet gender might be a salient concern in a country with demonstrated bias in assessment of teachers by head teachers (Beg, Fitzpatrick, and Lucas 2021). In column 3 of Table A10 we test for heterogeneity based on student gender by interacting each of the four treatment variables with female. In all cases, the interaction term is positive and in three cases statistically significant–female test scores increased more than male test scores in the three interventions that had a remedial or differentiated focus. In all cases we reject that the coefficient on the sum of the interaction and the main effect equals 0 indicating female students scores statistically significantly increased across all interventions. The main effects, i.e. the effect on male students, are now statistically insignificant for both the assistant-split and the teacher-led differentiated instruction model. As teachers were 10 percentage points more likely to be women than the assistants, the additional improvement for girls in the assistant-led arms is not likely due to gender-matching role model effects.

7 Discussion

7.1 Mechanisms

Recall the relative comparisons from Section 3.2 and Figure 1. Based on the point estimates when students were in grade 3 (follow-up 2), once remedial instruction is included, a smaller class size has almost the same point value as an extra instructional hour (T1 vs T2, 0.007SD difference) potentially because the status quo in-class work was at a wholly inappropriate learning level for students who needed remedial instruction. Confirming that grade-level material was too difficult, an assistant focusing on remedial instruction instead of grade level

content during school increased scores by an additional 57 percent (T2 vs T3). Finally, more homogeneous classrooms led by civil service teachers were a weak substitute for remedial-focused assistants as score increases were only 54 percent of the size in the teacher-led differentiated instruction arm. Each intervention had fidelity of implementation issues, which we discuss more below, yet these relative effect sizes are what one would expect from a similarly supported, fully government implemented program.

7.2 Implementation Challenges

Despite positive effect sizes, all four interventions faced implementation difficulties related to the challenges in scaling a program within existing government structures and none were exactly implemented as intended. In Panel A of Appendix Table A11 we repeat the effect size of each intervention and basic metrics along the causal chain of implementation–the educator attending training, being present at school, using the intervention materials, and adhering to the intervention. All interventions had over 85 percent adherence to training, but our educators' attendance and use of intervention materials were low.

Three specific challenges related to working within existing systems likely contributed to the lower levels of adherence. First, the program used existing production and distribution systems for material delivery, leading to delays into the second academic year. Second, assistants' payments were delayed, including a span of 8 months with no payment. Over the three years of spot-checks, of those assistants who were present to be surveyed, 60 percent reported delayed salaries as one of their main challenges. Further, schools reported assistant strikes due to non-payment of salaries. Third, teachers and assistants were only subject to existing support and incentives. Assistants likely had stronger incentives than teachers since this was a multiple year intervention, it was the primary source of income for about 70 percent of assistants, and they could be fired by their schools, but most did not aspire to teach in the future, in contrast to the contract teachers in Duflo, Dupas, and Kremer (2015).

These three deficiencies show areas in which stronger government systems could increase

the likelihood of positive effects when scaling previously effective interventions and potentially explain why Vivalt (2020) found programs that were implemented by academics or NGOs had larger effects that those that were government-implemented.

7.3 Comparison to Other Studies

Figure 3 plots our effect sizes (solid bars) relative to previous interventions that are the most similar to ours–class size reductions, tracking, assistant-led remedial instruction, and teacher-led differentiated instruction. Based on existing evidence, whether our four interventions could be effective without NGO support was unclear as can be seen by the small and statistically insignificant effect sizes of the only prior government intervention (Bold et al. 2018) and the two teacher-led differentiated instruction versions in India without an extra supervisory layer (Banerjee et al. 2017). Further, the assistant-led interventions had not previously been tested in an exclusively government implemented program.

Relative to the previous attempt at embedding a class size reduction in an existing government system (red and black downward diagonals), we found larger and statistically significant effects (Bold et al. 2018).[5] Teacher only interventions (blue bars) had previously had small, statistically insignificant point values until an NGO created supervisory layer was implemented along with an extra hour of school (Banerjee et al. 2017). Appendix Table A11 contains the same steps on the causal chain as those that appear for this study. The adherence to training was among the highest in this study but the fidelity of implementation was among the lowest–between the unsuccessful teacher-based interventions in India and the more successful contract teacher interventions in Kenya despite repeated refresher trainings throughout the intervention years.

The smaller effect sizes relative to some NGO implementations were likely not do to lack of initial political will or interest. This intervention started as a government program

[5]Muralidharan and Sundararaman (2013) evaluated a partnership between an Indian NGO, Azim Premji Foundation, and the government of Andhra Pradesh, India that provided contract teachers to schools. Schools could use them as they wished and in many cases used them to reduce the number of multi-grade classrooms, a different intervention.

in contrast to Banerjee et al. (2017) and Bold et al. (2018) who studied the government implementation of NGO programs. Instead, fidelity of implementation by the front line civil servants was lower than under NGO implementation and waning adherence was likely due to the existing weak supervisory and payment systems.

8 Cost Effectiveness

We present an overstatement of the costs of the program as implemented–providing the costs of the program using the ingredients method as the program was designed, including the on time delivery of materials and assistant payments. Based on estimates of scaling a single program to the entire country, the per student annual costs would be $19.60 for the remedial assistant interventions, $18.77 for the assistant-split, and $10.65 for teacher-led targeted instruction. When considering cost-effectiveness, we follow Kremer et al. (2013) and put each intervention on an effect size per $100 scale based on the point estimate of the effect. Our students received effectively two years of the intervention, spread across three academic years. The effect sizes per $100 are 0.21SD for the assistant split, 0.36SD for the teacher-led differentiated instruction, and 0.38SD for the assistant-led remedial during or after school. The similarity of the point estimates of the three differentiated or remedial interventions are remarkable–the assistant-led ones cost approximately twice as much per student with approximately twice the benefit.[6]

Because we have a multiple year intervention considering the cost effectiveness of a shorter duration, i.e. lower dosage, of the program is tempting. The first follow-up occurred near the start of the second school year, about two terms into the program. As the effect size for the combined English and Math score at this first follow-up is between 53 and 69 percent of the point values for the second follow-up and the costs were less than half, a shorter dose of the program appears to be more cost effective, yet the program deteriorated over time with increased educator absenteeism and decreased fidelity of implementation. Therefore,

[6] As is common in the literature, this evaluates each study at its coefficient point value.

whether the implementation of a model closer to the ideal between the two follow-ups would yield larger effect sizes is unknown.

9 Conclusions

Many countries that have eliminated the barriers to schooling are now beset with the dual challenge of heterogeneous classrooms with low average levels of learning. We used a 500 school RCT to test four government-designed interventions to improve student achievement in lower primary school across 42 districts in all 10 regions in Ghana. Three versions used an existing government program to hire assistants, primarily from the local community, to act as assistants. The assistants either operated a remedial pull-out program, provided after school remedial lessons, or divided the learners between the teacher and themselves for part of the school day. The final intervention used existing teachers who were instructed to divide three grade-levels of students by learning level instead of grade-level for a part of each day.

All four interventions increased student learning based on test administered at the end of grade 3 for those students who started the program in grade 1. Effect sizes range from 0.08 (teacher-led differentiated instruction) to 0.15 (assistant-led remedial after school) standard deviations. Students who were exposed to the program starting in grade 2 and tested at the end of grade 4, one year after ending the program have smaller effect sizes (with the exception of the assistant split) and the teacher-led model is no longer statistically significant. We find no evidence that the program affected the non-cognitive outcomes of student attendance, drop-out, or likelihood of being demoted. The teacher-led model increased the likelihood that teachers were engaged with students and using teaching and learning materials. When considering cost effectiveness, the assistant-led after school remedial program, assistant-led during school remedial program, and teacher-led differentiated instruction program were similarly cost effective–the effect sizes and costs of the first two were approximately twice the size of the third.

All models faced issues of material delays, teacher and assistant absenteeism, and weak

mechanisms for support and monitoring, factors that could potentially be remedied with additional training and support for managerial layers of the civil service. Stronger adherence to the prescribed model could result in larger effect sizes.

References

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy 2*(1), 1–30.

Beg, S., A. Fitzpatrick, and A. M. Lucas (2021, May). Gender bias in assessments of teacher performance. *AEA Papers and Proceedings*.

Beg, S. A., A. M. Lucas, W. Halim, and U. Saif (2019, March). Engaging teachers with technology increased achievement, bypassing teachers did not. Working Paper 25704, National Bureau of Economic Research.

Bold, T., M. Kimenyi, G. Mwabu, A. Ngángá, and J. Sandefur (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics 168*, 1 – 20.

Duflo, E., P. Dupas, and M. Kremer (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics 123*, 92–110.

Duflo, E., P. Dupas, and M. Kremera (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review 101*(5), 1739–1774.

Duflo, E., R. Hanna, and S. P. Ryan (2012). Incentives work: Getting teachers to come to school. *American Economic Review 102*(4), 1241–78.

Evans, D. K. and A. M. Acosta (2020, August). Education in africa: What are we learning. Working Paper 542, Center for Global Development.

Evans, D. K. and F. Yuan (2019, July). What we learn about girls education from interventions that do not focus on girls. Working Paper 513, Center for Global Development.

Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal (forthcoming). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources*.

Hartwell, A. (2010). National literacy acceleration program (nalap) implementation study. Working paper, Education Quality for All Project (EQUALL).

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Kremer, M., B. Conner, and R. Glennerster (2013). The challenge of education and learning in the developing world. *ScienceMag 340*.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Lucas, A. M. and I. M. Mbiti (2012). Access, sorting, and achievement: The short-run effects of free primary education in kenya. *American Economic Journal: Applied Economics 4*(4), 226–53.

Ministry of Education (2014, May). Ghana 2013 national education assessment technical report. Technical report.

Muralidharan, K. and V. Sundararaman (2013, September). Contract teachers: Experimental evidence from india. Working Paper 19440, National Bureau of Economic Research.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*.

A Appendix

A.1 Data Collection and Test Design

In this sub-section we describe the data collection rounds and achievement test design. Figure 2 in the main text displays the academic year, intervention, and data collection timeline. In the Section A.2, we provide summary statistics showing baseline balance across the treatment arms.

Baseline

The baseline occurred October to December 2010, the first term of the 2010-2011 academic year. Head teachers, i.e. school principals, teachers, and grade 1 and 2 students were interviewed. We tested selected students using bespoke exams, see "Test Design" below for more details on the exams.

Spot-checks

Between the baseline and final achievement follow-up, we conducted six rounds of additional data collection through spot-checks, visiting a sub-sample of schools each time. These visits occurred once each term starting with the third term of the first year (June and July of 2011) and ending with the second term of the third year (January through April of 2013). Each round included classroom observations and recording the presence or absence of the head teacher, the teachers, the assistants, and the baseline students. Further, we asked the teachers the current grade level of each student and whether they were assigned to the remedial section (as relevant). For absent students we asked teachers whether the student was still attending the school.

Achievement Follow-ups

We conducted two rounds of full achievement follow-ups. The first was November and December of 2011, the end of the first term of year two, approximately one year after the baseline and the second term after implementation. The second was about 18 months later in June and July 2013, near the end of the third academic year, two full academic years after

the start of implementation.

In each follow-up we sought to interview the same students from baseline regardless of their grade at follow-up. In follow-up 1, our baseline grade 1 and 2 students should have been in grades 2 and 3. In follow-up 2, we focused on the same students, who should have been in grades 3 and 4. We tested grade 4 students (i.e. those who were in grade 2 at baseline) to study the intervention's effects one year after leaving the program.[7] The data collection strategy included testing students present at school, encouraging absent students to come to school for testing, and tracking those who did not come to school. The survey teams attempted to test all baseline students.

Test Design

We developed the bespoke exams in collaboration with the Assessment Services Unit of the Curriculum Research and Development Division of GES. We further had the support of a psychometrician and piloted the exams on 300 students to test for validity and reliability.

The tests contained critical objectives of the official curriculum for grades 1-3, covering a range of skills, beyond what the lowest level differentiated instruction or remedial materials covered and beyond what is contained in the Annual Status of Education Report (ASER), the Early Grade Reading Assessment (EGRA), or the Early Grade Mathematics Assessment (EGMA) exams. Students were tested in English, math, and the school's NALAP language. One common test was used to test pupils across all grades. Questions were not ordered by difficulty and students were asked all questions even if they had incorrectly answered previous questions.

Exams in each round where similar in spirit but contained different questions. The baseline exam was entirely oral. The follow-up exams included additional written grade-level specific components based on the students' expected grades, i.e. grade 1 baseline students had additional grade 2 written content in follow-up 1 and grade 3 content in follow-up 2. Within round achievement is converted to latent scores using item response theory and standardized using the control group mean and standard deviation pooled across both grade

[7]Due to budgetary reasons in follow-up 2 we only tested a sub-sample of the original grade 2 sample.

1 and grade 2 students at baseline to ensure comparability across all achievement tables.

A.2 Summary Statistics and Baseline Balance

Appendix Tables A1 through A3 provide summary statistics and show baseline balance across students (Appendix Table A1), teachers and schools (Appendix Table A2), and assistants (Appendix Table A3). In Tables A1 and A2, columns 1 through 5 contain the means by treatment status as indicated at the top of the column. Column 6 contains the F-test and p-value for a test for the equality across all five columns. Table A3 compares the three arms with assistants with means in columns 1 through 3 and the F-test and p-value in column 4. Across all three tables, we fail to reject the null hypothesis that the means are equal in all but one case. We find the likelihood of assistants living in the community prior to the intervention is statistically different.

Students

[Appendix Table A1 about here]

Despite an official age of entry of 6, the overall average age of students at baseline was about 8.5. Students were on average 7.8 in grade 1 and 9.0 in grade 2. Attesting to the expansion of primary school access, about half of the sample had a literate father and about a third had a literate mother.

Teachers

[Appendix Table A2 about here]

A few means of note for the teachers: Just over half of the teachers were female, and they were on average about 36 years old. About 60 percent lived in the community in which they taught and had on average about 10 years of experience as teachers. Around 85 percent were employed directly by Ghana Education Services (GES), indicating that they were permanent teachers. The other 15 percent were employed by NYEP, the National Service Secretariat

(NSS) as part of the year of required national service, the community, or an NGO or were unpaid volunteers.

Schools

Summary statistics and baseline balance checks collected at the school level appear in Panel B of Table A2. As with the teachers and students, we do not find statistically significant differences across the 5 arms. Across all three lower primary grades, the average total enrollment was about 119, or about 37 students per grade. On average about 3.5 teachers were assigned to these three grades, resulting in an average pupil-teacher ratio of 35 to 1. Grade 1 cohorts were on average the largest cohort (42 students) with the largest pupil-teacher ratio (36 students per teacher). About one quarter of schools had electricity. To provide additional context on the level of infrastructure, over 80 percent of schools had cement or concrete floors, a metal roof, and cement or concrete walls.

Assistants

[Appendix Table A3 about here]

Table A3 contains the demographic characteristics of assistants, collected during spot-check rounds because they had not yet been hired at the baseline. One concern when comparing the effects of the different arms could be that schools selected assistants differently based on the intervention, e.g. the characteristics of a during-school assistant might be different than an after-school assistant. According to the demographic data collected, assistants were statistically similar across the three treatment arms with one exception–after school assistants were more likely to be living in the community prior to being hired. On average assistants were 25 years old. About 40 percent were women and about half worked for income prior to being hired as an assistant. Upon being hired, about 70 percent reported that the intervention income was their main source of income. Almost 80 percent, more in the after-school arm, reported living in the community prior to being hired for the intervention. Over half had some teaching experience. This experience including tutoring, teaching in private schools, and teaching in government schools.

According to the instructions given to the communities, all assistants should have been interviewed, been asked to present evidence that they passed the high school exit exam, completed high school, and been able to read, write and speak the school's official local (NALAP) language. According to the assistants these instructions were largely followed. Based on self-reports almost three-quarters were interviewed, about 65 percent were asked about their exit exam scores or whether they passed the high school exit exam, and over 90 percent were able to read, write, and speak the NALAP language, were able to speak the students' most common primary language, and completed high school. Unlike the contract teachers in Duflo, Dupas, and Kremer (2012 and 2015) who were all aspirant teachers, only about 40 percent of these assistants aspired to teach in the future.

A.3 Additional Estimates and Figures

Table A4 provides subject-specific test scores outcomes for the cohort who was treated starting in grade 1, should have been grade 2 at follow-up 1, and grade 3 at follow-up 2. Panel A includes all test questions while Panel B focuses on the questions most similar to the ASER test in India. Only the assistant-led remedial during school statistically significantly increased the local language score when grade-level material is included. In contrast, all three models with a focus on remedial skills statistically significantly improved the foundational local language test scores. Given the delays and complications with the NALAP program, grade-level local language content could have been too difficult for all students.

[Table A4 about here]

Table A5 tests for persistent achievement effects for students who were in grade 2 at baseline, experienced the intervention in grades 2 and 3, and were in expected grade 4 at follow-up 2.

[Table A5 about here]

In Table A6 we test for differential selection into test taking by treatment status, finding none.

[Table A6 about here]

Despite finding no evidence of differential selection into test taking by treatment status, Table A7 provides Lee (2009) bounds for our estimates.

[Table A7 about here]

Table A8 shows that the intervention had no effect on students' non-cognitive outcomes–being present on an unannounced day, no longer attending the baseline school, or repeating a grade.

[Table A8 about here]

Table A9 reports heterogeneity by baseline test score (column 1), whether the student was in the top two thirds of the baseline by grade test score distribution (an approximation for not receiving remedial support (column 2)), and whether the student was female (column 3).

[Table A10 about here]

Table A10 provides the effect sizes from from this paper, rows denoted (1), and the other studies in Figure 3 along with the steps along the causal chain–educators being trained, being present, using the intervention materials, and adhering to the intervention.We thank James Berry for providing the combined math and literacy test scores for the Banerjee et al. (2016) set of interventions.

[Table A11 about here]

Figure 1: Intervention Components and Graphical Conceptual Framework

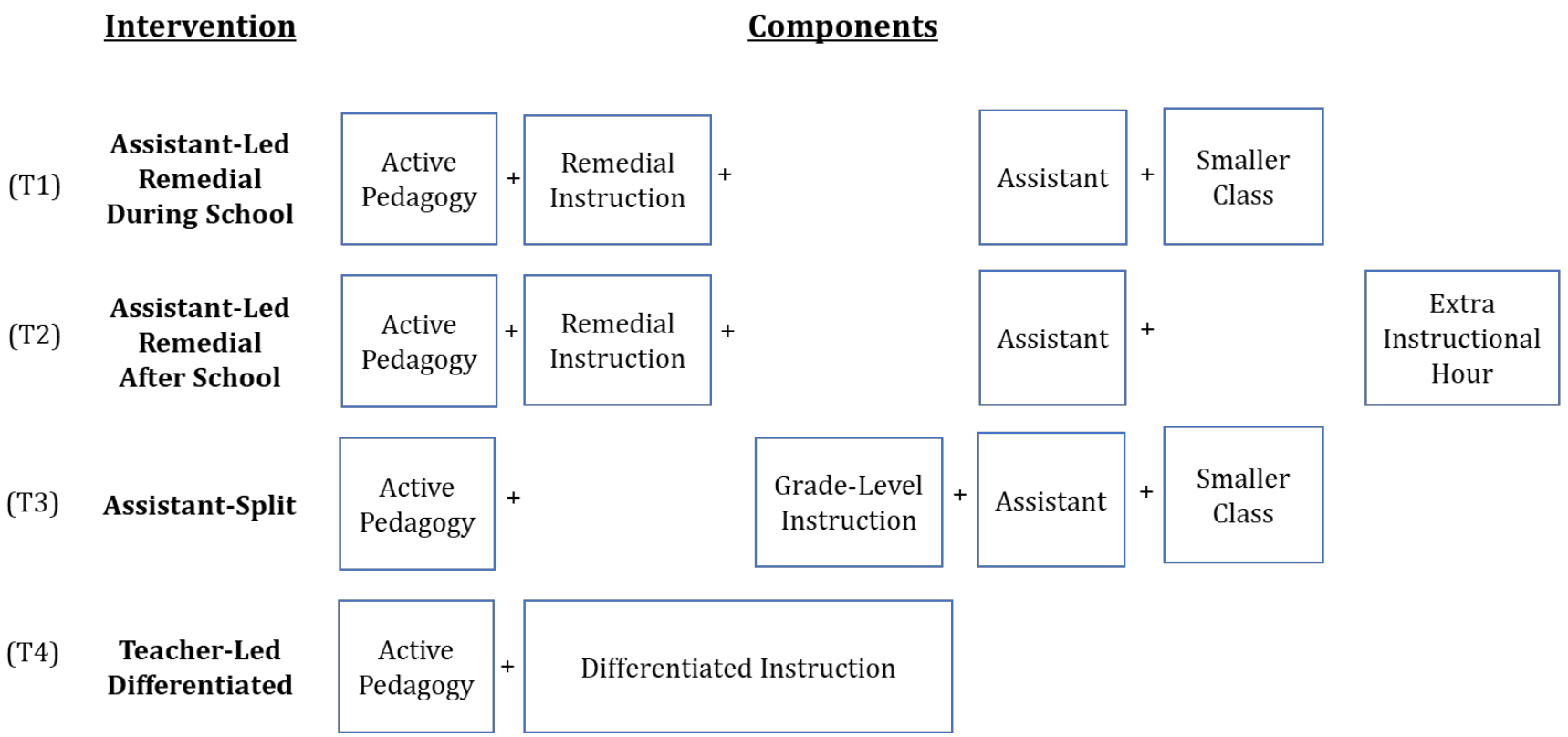
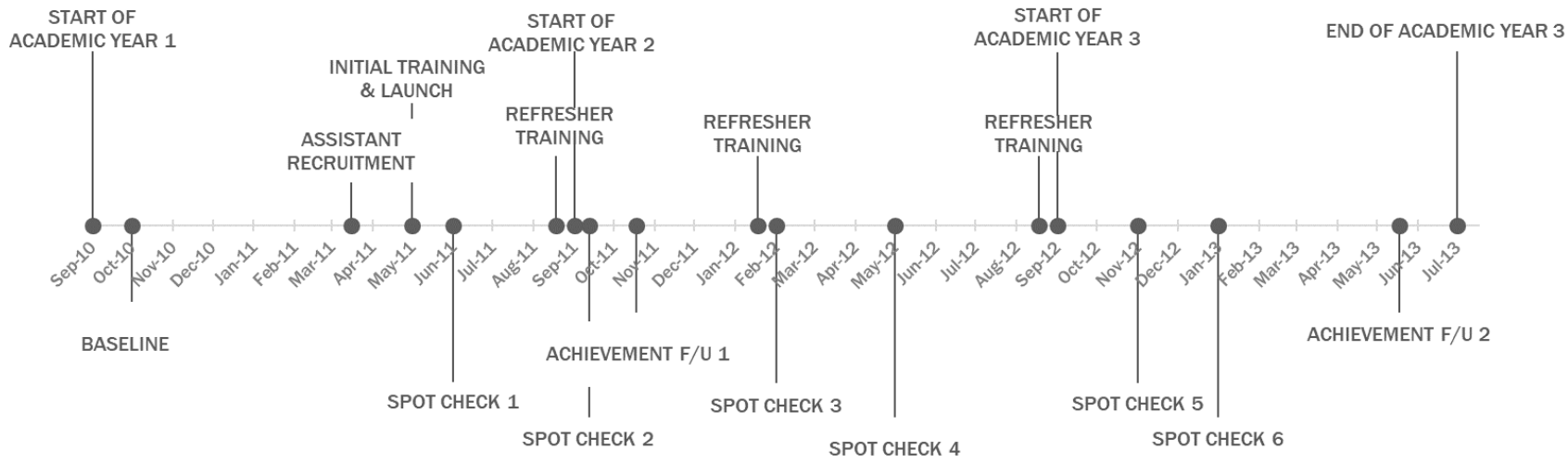
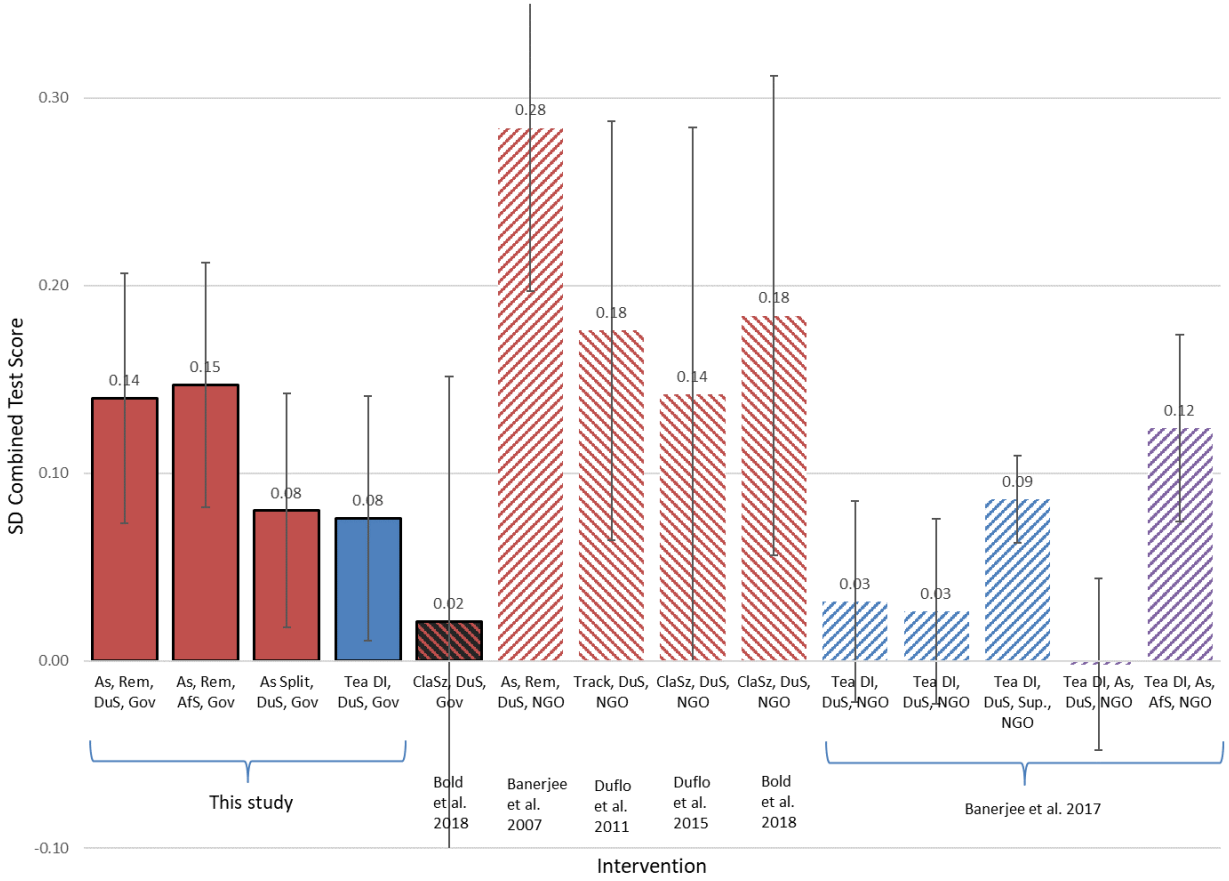


Figure 2: Academic Year, Implementation, and Data Collection Timeline



Notes: Labels above the line are academic year and implementation milestones. Those below the line are the nine data collection points.

Figure 3: Comparisons across Interventions and Contexts



Notes: Solid bars are this study. The color of the bar or diagonal represent the intervention educator: red are assistants or additional teachers who worked separately with students, blue are teachers only, purple are both assistants and teachers. The second color represents the implementer: Black diagonals and outlines are government and white diagonals are NGO. Downward sloping diagonals are in Kenya, upward sloping diagonals are in India. Error bars are 90% confidence intervals. The first four bars reproduce the two year effect sizes from Table 1. See Appendix Table A10 for sources from the other studies. As=Assistants. Rem=Remedial. DuS=during school. AfS=After school. DI=Differentiated Instruction. Sup=extra supervisory layer. Gov=government implemented. NGO=NGO implemented.

Table 1: Achievement in Math and English

		Exposed Grades 1-3				Exposed Grades 2-3, Tested in Grade 4	
		All Questions		Foundational Questions		All Questions	Grade 1-3 Questions
		Grade 2 (1)	Grade 3 (2)	Grade 2 (3)	Grade 3 (4)	(5)	(6)
(1)	Assistant Led Remedial, During School	0.096**	0.140***	0.120***	0.147***	0.071	0.102**
		(0.039)	(0.046)	(0.038)	(0.047)	(0.044)	(0.042)
(2)	Assistant Led Remedial, After School	0.101**	0.147***	0.139***	0.159***	0.084*	0.114***
		(0.042)	(0.045)	(0.039)	(0.045)	(0.045)	(0.043)
(3)	Assistant Split	0.042	0.080*	0.054	0.072*	0.116**	0.133***
		(0.041)	(0.043)	(0.037)	(0.043)	(0.047)	(0.046)
(4)	Teacher Led Differentiated Instruction	0.052	0.076*	0.092**	0.125***	0.013	0.052
		(0.043)	(0.045)	(0.040)	(0.043)	(0.047)	(0.046)
	P-value of Test of Equality						
	1 = 2	0.91	0.88	0.65	0.80	0.77	0.79
	1 = 3	0.19	0.20	0.08	0.12	0.33	0.49
	1 = 4	0.31	0.18	0.49	0.64	0.22	0.27
	2 = 3	0.19	0.15	0.03	0.06	0.50	0.68
	2 = 4	0.29	0.13	0.27	0.45	0.15	0.19
	3 = 4	0.81	0.92	0.34	0.23	0.04	0.10
	Observations	8,653	8,004	8,654	8,004	4,302	4,302
	R-squared	0.55	0.47	0.45	0.39	0.49	0.43

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Outcomes are standard deviation test score changes. Standard errors clustered at the school level appear in parenthesis. All regressions include baseline test scores and dummy variables for strata and female. Sample of students tested at baseline and relevant follow-up round. Columns 1-4: Students in grade 1 at baseline. Columns 5 and 6: Students in grade 2 at baseline.

Table 2: Unannounced Visits--Time on Task and Fidelity of Implementation

	Head Teacher Present	Teachers				Target Educator Teaching to a Group
		Present	In Classroom	Engaged with Students	Using Materials	
	(1)	(2)	(3)	(4)	(5)	(6)
Assistant Led Remedial, During School	0.135***	0.005	0.030	0.038	-0.002	0.333***
	(0.038)	(0.024)	(0.029)	(0.028)	(0.027)	(0.027)
Assistant Led Remedial, After School	0.058	0.007	0.023	0.027	-0.009	0.412***
	(0.039)	(0.025)	(0.029)	(0.026)	(0.030)	(0.031)
Assistant Split	0.018	-0.029	-0.003	0.025	-0.011	0.268***
	(0.039)	(0.027)	(0.029)	(0.026)	(0.027)	(0.027)
Teacher Led Differentiated Instruction	0.089**	0.030	0.052*	0.110***	0.206***	0.058***
	(0.039)	(0.024)	(0.029)	(0.027)	(0.031)	(0.015)
Test of Equality						
F-Statistic	3.36	1.66	1.17	4.68	19.49	6.59
p-value	0.02	0.18	0.32	0.00	0.00	0.00
Observations	1,892	6,324	6,305	6,143	2,378	1,793
R-squared	0.07	0.07	0.09	0.08	0.17	0.21
Control Group Mean	0.51	0.69	0.52	0.36	0.05	0.00

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Columns 3 and 4: not conditional on being present. Column 5: conditional on a teacher being present. Column 6: At the school level. Whether group meetings were being held in the assistant intervention or teachers were either working with a subset of their students or had divided students by learning level across classrooms.

Table 3: Unannounced Visits--Fidelity of Implementation Over Time

	Target Educator Teaching to a Group		
	Academic Year 1 (1)	Academic Year 2 (2)	Academic Year 3 (3)
Assistant Led Remedial, During School	0.501*** (0.054)	0.313*** (0.041)	0.204*** (0.042)
Assistant Led Remedial, After School	0.554*** (0.054)	0.383*** (0.042)	0.342*** (0.052)
Assistant Split	0.374*** (0.053)	0.296*** (0.037)	0.102*** (0.034)
Teacher Led Differentiated Instruction	0.042 (0.033)	0.055** (0.023)	0.061** (0.029)
Test of Equality			
F-Statistic	3.20	1.42	8.41
p-value	0.04	0.24	0.00
Observations	437	902	454
R-squared	0.32	0.22	0.20
Control Group Mean	0.00	0.00	0.00

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. At the school level. Whether group meetings were being held in the assistant intervention or teachers were either working with a subset of their students or had divided students by learning level across classrooms.

Appendix Table A1: Summary Statistics--Students

	Treatment				Control	Test of Equality F-stat (p-value)
	Assistant Led Remedial Instruction		Assistant Only	Teacher Led Differentiated Instruction		
	During School	After School				
	(1)	(2)	(3)	(4)	(5)	(6)
Combined English and Math Test Score	0.000	-0.009	0.012	-0.078	0.000	0.43
	(0.95)	(1.01)	(0.99)	(0.93)	(1.00)	(0.79)
English Test Score	-0.003	-0.001	0.004	-0.091	0.000	0.53
	(0.97)	(1.03)	(1.02)	(0.94)	(1.00)	(0.72)
Math Test Score	0.003	-0.015	0.017	-0.051	0.000	0.28
	(0.95)	(0.99)	(0.97)	(0.94)	(1.00)	(0.89)
Local Language Test Score	-0.019	-0.044	-0.007	-0.071	0.000	0.22
	(0.98)	(1.02)	(0.93)	(0.95)	(1.00)	(0.93)
Student wore a clean, good quality uniform	0.28	0.27	0.30	0.23	0.27	1.62
	(0.45)	(0.45)	(0.46)	(0.42)	(0.45)	(0.17)
Student had shoes	0.90	0.91	0.91	0.89	0.91	0.23
	(0.30)	(0.29)	(0.29)	(0.31)	(0.28)	(0.92)
Student Age	8.37	8.34	8.52	8.48	8.49	0.00
	(0.30)	(0.29)	(0.29)	(0.31)	(0.28)	(0.92)
Father Literate	0.46	0.48	0.46	0.45	0.47	0.25
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.91)
Mother Literate	0.31	0.36	0.34	0.34	0.33	0.68
	(0.46)	(0.48)	(0.47)	(0.47)	(0.47)	(0.61)
Someone at home helps with homework	0.64	0.62	0.66	0.62	0.64	1.01
	(0.48)	(0.49)	(0.47)	(0.48)	(0.48)	(0.40)
Self-reported absences in the last week	0.76	0.80	0.86	0.84	0.81	0.60
	(1.30)	(1.37)	(1.41)	(1.39)	(1.35)	(0.66)

Notes: Columns (1) - (5): Standard deviations appear in parenthesis. Column (6): F-statistic with p-value in parenthesis of a test of equality across all treatment arms, taking into account clustering by school. Averages calculated over grade 1 and 2 baseline students who completed an examination. Test scores standardized by subject with control mean of 0, standard deviation of 1.

Appendix Table A2: Summary Statistics--Teachers and Schools

	Treatment				Control	Test of Equality F-stat (p-value)
	Assistant Led Remedial Instruction		Assistant Only	Teacher Led Differentiated Instruction		
	During School	After School				
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teachers						
Female	0.55	0.56	0.50	0.51	0.53	0.46
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.76)
Age	35.85	36.68	35.38	34.82	34.73	1.14
	(12.19)	(11.30)	(10.77)	(11.44)	(11.61)	(0.34)
Live in the community	0.60	0.57	0.63	0.62	0.64	0.65
	(0.49)	(0.50)	(0.48)	(0.49)	(0.48)	(0.63)
Years of Experience	10.17	11.06	9.85	9.67	9.83	0.75
	(10.12)	(9.85)	(9.67)	(9.85)	(10.31)	(0.56)
Employed by Ghana Education Services	0.82	0.86	0.84	0.86	0.86	0.50
	(0.38)	(0.35)	(0.37)	(0.35)	(0.34)	(0.74)
Panel B: Schools						
Lower Primary Enrollment	117.0	120.0	121.1	118.8	118.8	0.06
	(62.4)	(60.8)	(63.7)	(74.9)	(94.5)	(0.99)
Number of Lower Primary Teachers	3.52	3.44	3.49	3.57	3.28	0.69
	(1.50)	(1.10)	(1.44)	(1.53)	(1.25)	(0.60)
Lower Primary Pupil Teacher Ratio	35.5	35.0	36.2	33.8	35.1	0.26
	(20.6)	(14.6)	(17.8)	(17.5)	(19.0)	(0.90)
Electricity	0.33	0.23	0.28	0.23	0.23	0.97
	(0.47)	(0.42)	(0.45)	(0.42)	(0.42)	(0.43)

Notes: Columns (1) - (5): Standard deviations appear in parenthesis. Column (6): F-statistic with p-value in parenthesis of a test of equality across all treatment arms, taking into account clustering by school. Panel A: Averages calculated over all teachers who completed a baseline survey. Panel B: Averages calculated over all 500 study schools.

Appendix Table A3: Summary Statistics--Assistant Characteristics

	Treatment			Test of Equality F-stat (p-value)
	Assistant Led Remedial Instruction		Assistant Only	
	During School	After School		
	(1)	(2)	(3)	(4)
Age	24.98 (4.86)	25.12 (5.37)	25.14 (4.73)	0.05 (0.95)
Female	0.48 (0.50)	0.40 (0.49)	0.41 (0.49)	1.19 (0.31)
Any Income Pre-Intervention	0.58 (0.49)	0.52 (0.50)	0.58 (0.50)	0.77 (0.46)
Main Income Now from Intervention	0.74 (0.44)	0.68 (0.47)	0.69 (0.46)	0.60 (0.55)
Lived in Community Pre-Intervention	0.76 (0.43)	0.86 (0.35)	0.77 (0.42)	3.06 (0.05)
Teaching Experience	0.62 (0.49)	0.59 (0.49)	0.56 (0.50)	0.60 (0.55)
Interviewed	0.75 (0.43)	0.73 (0.44)	0.68 (0.47)	0.76 (0.47)
Asked about Scores or Passing	0.70 (0.46)	0.67 (0.47)	0.61 (0.49)	1.09 (0.34)
Read, Write, and Speak Offical Local Language	0.91 (0.29)	0.92 (0.28)	0.91 (0.29)	0.08 (0.93)
Speak Most Common Student Primary Lanage	0.95 (0.22)	0.96 (0.20)	0.95 (0.21)	0.05 (0.95)
Completed High School	0.94 (0.23)	0.93 (0.25)	0.96 (0.19)	1.07 (0.35)
Aspire to Teach in the Future	0.40 (0.49)	0.35 (0.48)	0.39 (0.49)	0.60 (0.55)

Notes: Columns (1) - (3): Standard deviations appear in parenthesis. Column (4): F-statistic with p-value in parenthesis of a test of equality across all treatment arms, taking into account clustering by school.

Appendix Table A4: Subject-Specific Achievement

		English		Math		Local Language	
		Grade 2	Grade 3	Grade 2	Grade 3	Grade 2	Grade 3
		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Including Grade-Level Content							
(1)	Assistant Led Remedial, During School	0.081*	0.129***	0.093**	0.133***	0.105**	0.118**
		(0.043)	(0.048)	(0.039)	(0.045)	(0.052)	(0.057)
(2)	Assistant Led Remedial, After School	0.081*	0.174***	0.102**	0.100**	0.066	0.076
		(0.046)	(0.048)	(0.040)	(0.043)	(0.056)	(0.060)
(3)	Assistant Split	0.030	0.079*	0.046	0.070*	0.054	0.052
		(0.044)	(0.045)	(0.041)	(0.042)	(0.052)	(0.057)
(4)	Teacher Led Differentiated Instruction	0.025	0.085*	0.070*	0.056	0.076	0.060
		(0.045)	(0.047)	(0.041)	(0.043)	(0.053)	(0.061)
	P-value of Test of Equality						
	1 = 2	1.00	0.36	0.82	0.50	0.50	0.50
	1 = 3	0.26	0.29	0.25	0.18	0.33	0.26
	1 = 4	0.22	0.36	0.58	0.11	0.60	0.36
	2 = 3	0.29	0.05	0.18	0.51	0.84	0.70
	2 = 4	0.26	0.07	0.45	0.34	0.86	0.80
	3 = 4	0.91	0.90	0.57	0.75	0.68	0.90
	Observations	8,653	8,004	8,653	8,004	8,653	8,002
	R-squared	0.50	0.46	0.46	0.39	0.40	0.39
Panel B: Foundational Content Only							
(1)	Assistant Led Remedial, During School	0.111***	0.127**	0.105***	0.141***	0.119**	0.168***
		(0.042)	(0.050)	(0.037)	(0.044)	(0.049)	(0.056)
(2)	Assistant Led Remedial, After School	0.155***	0.178***	0.094**	0.109**	0.163***	0.176***
		(0.042)	(0.048)	(0.039)	(0.043)	(0.050)	(0.056)
(3)	Assistant Split	0.040	0.077*	0.057	0.053	0.092**	0.069
		(0.039)	(0.045)	(0.038)	(0.041)	(0.046)	(0.050)
(4)	Teacher Led Differentiated Instruction	0.096**	0.130***	0.069*	0.095**	0.085*	0.138***
		(0.042)	(0.044)	(0.041)	(0.042)	(0.046)	(0.053)
	P-value of Test of Equality						
	1 = 2	0.33	0.32	0.78	0.51	0.40	0.90
	1 = 3	0.10	0.30	0.19	0.06	0.58	0.07
	1 = 4	0.74	0.95	0.36	0.33	0.48	0.59
	2 = 3	0.01	0.03	0.33	0.21	0.16	0.05
	2 = 4	0.19	0.30	0.54	0.76	0.12	0.51
	3 = 4	0.18	0.22	0.76	0.34	0.87	0.19
	Observations	8,654	8,004	8,654	8,004	8,654	8,002
	R-squared	0.33	0.34	0.41	0.32	0.24	0.28

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Outcomes are standard deviation test score changes. Standard errors clustered at the school level appear in parenthesis. All regressions include baseline test scores and dummy variables for strata and female. Sample of students tested at baseline and relevant follow-up round. Columns 5 and 6: Local language is the official NALAP language, which might not be the primary language spoken by all (or most) students.

Appendix Table A5: Subject Specific Persistent Achievement

		English		Math	
		All Questions	Grade 1-3 Questions	All Questions	Grade 1-3 Questions
		(1)	(2)	(3)	(4)
(1)	Assistant Led Remedial, During School	0.033	0.063	0.100**	0.127***
		(0.049)	(0.047)	(0.044)	(0.044)
(2)	Assistant Led Remedial, After School	0.089*	0.114**	0.069	0.097**
		(0.049)	(0.046)	(0.045)	(0.045)
(3)	Assistant Split	0.088*	0.104**	0.131***	0.144***
		(0.050)	(0.047)	(0.047)	(0.047)
(4)	Teacher Led Differentiated Instruction	-0.020	0.021	0.046	0.077
		(0.050)	(0.047)	(0.047)	(0.048)
	P-value of Test of Equality				
	1 = 2	0.27	0.31	0.49	0.50
	1 = 3	0.27	0.41	0.50	0.70
	1 = 4	0.29	0.39	0.24	0.27
	2 = 3	0.99	0.85	0.19	0.31
	2 = 4	0.04	0.06	0.64	0.66
	3 = 4	0.04	0.10	0.09	0.17
	Observations	4,302	4,302	4,302	4,302
	R-squared	0.44	0.38	0.43	0.36

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Sample: students in grade 2 at baseline, exposed during grades 2 and 3, and tested again in follow-up 2 during expected grade 4, one year after exiting the program. All regressions include baseline test scores and dummy variables for strata and whether the student was female.

Appendix Table A6: Selection Into Test Taking

	Test Taker at			
	Follow-up 1		Follow-up 2	
	(1)	(2)	(3)	(4)
Assistant Led Remedial, During School	0.024	0.027	0.014	0.004
	(0.018)	(0.019)	(0.018)	(0.018)
Assistant Led Remedial, During School X BL Score		0.004		-0.017
		(0.023)		(0.026)
Assistant Led Remedial, After School	0.013	0.003	0.000	-0.016
	(0.017)	(0.017)	(0.017)	(0.020)
Assistant Led Remedial, After School X BL Score		-0.017		-0.030
		(0.022)		(0.024)
Assistant Split	0.028	0.015	0.007	-0.007
	(0.017)	(0.018)	(0.017)	(0.020)
Assistant Split X BL Score		-0.023		-0.026
		(0.022)		(0.024)
Teacher Led Differentiated Instruction	-0.023	-0.021	-0.012	-0.029
	(0.021)	(0.022)	(0.018)	(0.020)
Teacher Led Differentiated X BL Score		0.003		-0.030
		(0.029)		(0.025)
Test of Jointly Equal 0				
F-Statistic	2.19	1.49	0.55	0.92
p-value	0.07	0.20	0.70	0.45
Observations	10,977	10,977	10,977	10,977
R-squared	0.03	0.03	0.01	0.01
Control Group Mean	0.78		0.73	

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. All regressions include baseline test scores and dummy variables for strata and female.

Appendix Table A7: Lee Bounds

	Combined English and Math		English		Math		Local Language	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Follow-up 1								
Assistant Led Remedial, During School	0.105***	0.110***	0.082*	0.087**	0.112***	0.116***	0.093*	0.107**
	(0.040)	(0.039)	(0.044)	(0.044)	(0.039)	(0.037)	(0.052)	(0.050)
Assistant Led Remedial, After School	0.113***	0.112***	0.095**	0.097**	0.112***	0.109***	0.069	0.086
	(0.042)	(0.042)	(0.046)	(0.046)	(0.041)	(0.041)	(0.056)	(0.054)
Assistant Split	0.048	0.061	0.026	0.039	0.062	0.073*	0.065	0.063
	(0.040)	(0.040)	(0.042)	(0.043)	(0.040)	(0.040)	(0.052)	(0.051)
Teacher Led Differentiated Instruction	0.060	0.060	0.029	0.029	0.081**	0.082*	0.080	0.079
	(0.042)	(0.042)	(0.045)	(0.045)	(0.041)	(0.042)	(0.053)	(0.053)
Test of Equality								
F-Statistic	1.17	0.98	1.12	1.00	0.77	0.61	0.11	0.27
p-value	0.32	0.40	0.34	0.39	0.51	0.61	0.96	0.85
Observations	8,339	8,340	8,339	8,340	8,339	8,340	8,339	8,340
R-squared	0.51	0.51	0.44	0.47	0.44	0.42	0.38	0.37
Panel B: Follow-up 2								
Assistant Led Remedial, During School	0.141***	0.156***	0.129**	0.143***	0.132***	0.147***	0.117**	0.118**
	(0.047)	(0.047)	(0.050)	(0.050)	(0.046)	(0.045)	(0.057)	(0.055)
Assistant Led Remedial, After School	0.151***	0.154***	0.171***	0.173***	0.108**	0.112**	0.077	0.084
	(0.046)	(0.046)	(0.050)	(0.050)	(0.044)	(0.044)	(0.060)	(0.058)
Assistant Split	0.065	0.075*	0.066	0.077	0.054	0.063	0.049	0.055
	(0.043)	(0.043)	(0.047)	(0.047)	(0.041)	(0.041)	(0.057)	(0.054)
Teacher Led Differentiated Instruction	0.088**	0.093**	0.090*	0.095**	0.073*	0.077*	0.056	0.058
	(0.044)	(0.045)	(0.047)	(0.048)	(0.042)	(0.043)	(0.061)	(0.061)
Test of Equality								
F-Statistic	1.54	1.61	1.82	1.64	1.08	1.34	0.53	0.53
p-value	0.20	0.19	0.14	0.18	0.36	0.26	0.66	0.66
Observations	7,854	7,853	7,854	7,853	7,854	7,853	7,853	7,853
R-squared	0.43	0.44	0.43	0.44	0.35	0.34	0.38	0.37

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. All regressions include baseline test scores and dummy variables for strata and female. Each column pair represents the estimated upper and lower bound of the treatment effect for the outcome indicated at the top of the column based on Lee (2009). Columns 7 and 8: The local language is the offical NALAP language and may not be the primary language spoken by (any or all) of the students.

Appendix Table A8: Unannounced Visits - Student Outcomes

	Student Was		
	Present	No Longer Attending	Repeating a Grade
	(1)	(2)	(3)
Assistant Led Remedial, During School	0.002	-0.004	-0.003
	(0.022)	(0.018)	(0.026)
Assistant Led Remedial, After School	-0.004	0.019	0.011
	(0.019)	(0.018)	(0.026)
Assistant Split	0.006	0.016	0.031
	(0.020)	(0.017)	(0.026)
Teacher Led Differentiated Instruction	-0.009	0.017	0.040
	(0.020)	(0.018)	(0.029)
Test of Equality			
F-Statistic	0.23	0.69	0.96
p-value	0.88	0.56	0.41
Observations	10,944	10,767	9,141
R-squared	0.05	0.03	0.10
Control Group Mean	0.64	0.22	0.18

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school appear in parenthesis. Sample: grade 1 students at baseline. Not all rounds included all questions nor did teachers necessarily respond to all questions, thus the sample changes across columns.

Appendix Table A9: Heterogeneity in Achievement

		By Baseline Achievement		By Gender
		Test Score	Top Two Thirds of School by Grade	
		(1)	(2)	(3)
(1)	Assistant Led Remedial, During School	0.153***	0.118*	0.088*
		(0.048)	(0.064)	(0.050)
(2)	Assistant Led Remedial, During School X BL Ability	0.036	0.030	
		(0.047)	(0.063)	
(2)	Assistant Led Remedial, During School X Female			0.106**
				(0.053)
(3)	Assistant Led Remedial, After School	0.142***	0.105*	0.100*
		(0.051)	(0.058)	(0.051)
(4)	Assistant Led Remedial, After School X BL Ability	-0.018	0.059	
		(0.050)	(0.059)	
(4)	Assistant Led Remedial, After School X Female			0.097*
				(0.056)
(5)	Assistant Split	0.088*	0.044	0.062
		(0.047)	(0.056)	(0.050)
(6)	Assistant Split X BL Ability	0.022	0.052	
		(0.047)	(0.055)	
(6)	Assistant Split X Female			0.038
				(0.055)
(7)	Teacher Led Targeted Instruction	0.064	0.088	0.006
		(0.052)	(0.057)	(0.050)
(8)	Teacher Led Targeted X BL Ability	-0.046	-0.019	
		(0.050)	(0.059)	
(8)	Teacher Led Targeted X Female			0.145**
				(0.058)
	BL Ability or Female (varies by column)	0.537***	0.001	-0.021
		(0.034)	(0.045)	(0.040)
	P-value of Test of Coefficients Sum to 0			
	1+2=0		0.00	0.00
	3+4=0		0.00	0.00
	5+6=0		0.04	0.06
	7+8=0		0.16	0.01
	Observations	8,004	8,004	8,004
	R-squared	0.46	0.47	0.47

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Additional control variables: baseline test scores and dummy variables for strata and female. Column 1: BL ability is combined baseline English and math score. Column 2: BL ability is whether the student was in the top two thirds of the within school by grade baseline score distribution.

Appendix Table A10: Other Studies--Effect Sizes and Compliance

Intervention	Location	Effect on Student Test Scores	Educator				Duration (years)
			Trained	Present	Using Intervention Materials	Adhering to Intervention	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Government Implementation							
(1) Assistant, Remedial, During School	Ghana	0.140*** (0.046)	0.99	0.58	0.38	0.33	2
(1) Assistant, Remedial, After School	Ghana	0.147*** (0.045)	0.98	0.48	0.34	0.41	2
(1) Assistant, Partial Day Split	Ghana	0.080* (0.043)	0.97	0.63	0.39	0.27	2
(1) Teacher, Differentiated, During School	Ghana	0.076* (0.045)	0.85	0.73	0.11	0.06	2
(2) Contract Teacher, Full Year Split	Kenya	0.021 (0.090)	N/A	0.73	N/A	0.71	1.5
Panel B: NGO Implementation							
(3) Assistant, Remedial, During School	Mumbai + Vadodara, India	0.284** (0.060)					2
(4) Assistant, Remedial, After School+Village Education Committee Training	Jaunpur district, India	(a)					1
(5) Teacher, Differentiated, During School	Bihar state, India	0.0316 (0.0369)	0.67		0.58	0.04	1
(5) Teacher, Differentiated, During School	Uttarakhand state, India	0.0264 (0.0340)	0.28		0.26	0.10	1
(5) Teacher, Differentiated, During School+extra supervisory layer+devoted new school hour	Haryana state, India	0.0862*** (0.0161)	0.96		0.74	0.92	1
(5) Teacher+Assistant, Differentiated, During School	Uttarakhand state, India	-0.00195 (0.0315)	0.45		0.34	0.06	1
(5) Teacher, Differentiated, During School+Assistant, Remedial, After School	Bihar state, India	0.124*** (0.0344)	0.67		0.64		1
(6) Contract Teacher, Full Year Tracking	Western Province, Kenya	0.176** (0.077)	N/A	0.85	N/A	0.99	1.5
(7) Contract Teacher, Full Year Split	Western Province, Kenya	0.142 (0.098)	N/A	0.84	N/A		1.5
(2) Contract Teacher, Full Year Split	Kenya	0.184** (0.088)	N/A	0.63	N/A	0.69	1.5

Row Notes: (1) this study. (2) Bold et al. (2018), Table 4, columns 1 and 5; Table 9, Panel B, columns 1 and 2. (3) Banerjee et al. 2007. (4) Banerjee et al. 2010. (5) Banerjee et al. 2016, Table 3, Panel A, Panel B TM and TMV, Panel C TM, Panel D; Table 5, Panel A TM, Panel B TM and TMV, Panel C TaRL; additional calculations by Berry. (6) Duflo et al. 2011, Table 1 Panel E; Table 2, Panel A, column 2; Table 6 column 1. (7) Duflo et al. 2015, Table 3, Panel A, column 1.
Letter Notes: (a) results reported as increase in probability a student could read letters of 0.017** (0.007).
Other Notes: N/A indicates that this was not a feature of the intervention. Missing indicates the statistic was not provided.