EXPERIMENTAL EVIDENCE ON ALTERNATIVE POLICIES TO INCREASE LEARNING AT SCALE

Annie Duflo
Jessica Kiessel
Adrienne Lucas

WORKING PAPER 27298

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 May 2022

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 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, Pace Phillips, and the entire IPA Ghana team 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 numerous seminar and conference participants. 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.

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 May 2022
JEL No. I21,I25,I28,J24,O15

ABSTRACT

We partnered with the Ghanaian government to test simultaneously four methods of increasing achievement in schools with low and heterogeneous student achievement| assistant led remedial pull-out lessons, assistant led remedial after school lessons, assistant led smaller class sizes, or teacher implemented partial day tracking. Despite implementation issues, the interventions increased student learning by about 0.1SD, about 0.4SD when adjusting for the imperfect implementation, with no effects on attendance, grade repetition, or drop-out. Test score increases were larger for girls 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 smaller randomized controlled trials, yet 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, nationwide, randomized controlled trial that reached over 80,000 students, we test four models that built 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 show their potential, relative effectiveness, and effectiveness over time when fully designed and implemented by existing government systems. Results from this study have influenced the implementation of programs to improve education in India and Africa with the potential to reach 1 billion students at scale.

The Teacher Community Assistant Initiative (TCAI) was a Ghana Ministry of Education program that implemented four interventions to increase student learning using existing schooling and youth employment systems under the unifying theory that focusing more on individual learners could improve student outcomes. 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, i.e., pull-out remedial, 2) remedial learners outside of the school day. i.e., after school remedial, or 3) half of the classroom each day on grade-level content, i.e., classroom split.

The fourth intervention trained teachers to divide students into three learning levels for part of the day and focus instruction on students' learning levels, i.e., partial day tracking. 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 remediation can work at scale and existing systems can increase the amount of learning delivered. The interventions increased student learning by about 0.08 standard deviations (SD) after less than one year (point values 0.05SD to 0.11SD for each intervention) and 0.11SD after two years (point values 0.08SD to 0.15SD for each intervention) on tests that included grade level and foundational content, about 27 percent of a year of schooling in this context. We cannot statistically differentiate the four arms from each other when the exams include grade level content. When limiting the assessments to questions focused on foundational literacy and numeracy, the two remedial arms had a statistically larger effect than the classroom split. The interventions increased girls' test scores by about 0.1SD more than boys' scores with the differential gains concentrated in the interventions with the remedial or tracking component. The interventions did not affect students' likelihood of being present, dropping out, or repeating a grade level, common concerns with tracking and remedial programs. Test score increases persisted for students who were treated for about a year and tested one year after the end of the program.

As is common in government programs, implementation was imperfect: educators taught to their designated groups during only about one third of spot-check visits even though almost all had received training. That learning gains occurred despite low fidelity of implementation shows that focusing attention on specific learners, whether through smaller class sizes, tracking, or remedial lessons, is a robust strategy that confers learning gains even with incomplete adherence. Because not all students received the intended dosage, we estimate the treatment on the treated (TOT) using assignment to treatment at the school level as an instrument for the students being divided correctly during spot-checks. Using this instrumental variables approach, the interventions increased test scores by 0.3SD after less than one year and 0.4SD after two years.

In calculating costs, the partial day tracking was the least expensive as it relied on existing personnel while the assistant arms required assistant salaries. All four interventions had similar costs for trainings and materials. At the point values of the effect sizes, the cost-effectiveness is approximately the same for the pull-out remedial, after school remedial, and partial day tracking with worse cost-effectiveness for the classroom split. If the point values are equal, as could be the case given their statistical equivalence, then the partial day tracking is the most cost-effective.

Because the interventions shared common elements, we use a conceptual framework to show that if the point values are indeed equal, then a smaller class size, remedial instruction—whether as a pull-out program or an extra instructional hour—and tracking are almost perfect substitutes. If the focus is on foundational content where the effect sizes are statistically different, then these results show three important mechanisms: 1) remedial instruction is equally effective whether it's implemented as a pull-out or after school program, 2) a smaller class size focused on remedial instruction is more effective than one focused on grade level content, and 3) even though partial day tracking includes all learning levels, it increases average test scores no more than purely remedial instruction by assistants.

In addition to already influencing policy in both Africa and South Asia, our findings make three related contributions to the economics literature. First, our four alternatives to support teachers' transition to the new status quo incorporate four of the most promising findings from separate NGO-led interventions into a single study (Banerjee et al. 2017; Evans and Mendez Acosta 2021).[1] The issue of students not learning while in school has been highlighted as a primary concern in many countries, yet limited evidence exists on how to improve

[1] One potentially promising class of interventions we do not address are those using technology (see Beg et al. 2022 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).

learning at scale within existing government systems. All four of our interventions offer alternative ways to implement instruction more focused on individual learners—during or after school remedial lessons, by dividing the class in half, or having existing teachers specifically focus on a more homogeneous group of learners—building on Banerjee et al. (2007, 2010, 2017) and (Duflo et al. 2011).[2] By comparing the effects and cost effectiveness of the four alternatives together and 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. Tailored instruction increased learning yet implementation difficulties show that the capacity of the agency in charge of implementation might matter as much as the program design.

Second, this paper contributes to a broader literature on the importance of at-scale experiments implemented within existing systems. Most of the existing research on similar methods to increase student learning included at least one of the following: a highly motivated NGO, a researcher team heavily involved in implementation, a narrowly geographically selected sample, or additional personnel who were hired outside of normal government operations. This study relied on existing systems and included a randomly selected nationwide sample, features often lacking in experiments in development economics research (Muralidharan and Niehaus 2017). We show the potential for success of similar interventions at-scale and highlight the additional challenges of at-scale programs.

Third, we show that existing government structures have the capacity to increase learning in spite of rigid hierarchies and wages unrelated to productivity (Bau and Das 2020; Muralidharan et al. 2016). 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 version, government involvement started at the outset in

[2] The remedial pull-out intervention was inspired by the NGO-supported assistants in Banerjee et al. (2007) that increased learning in Mumbai and Vadodara cities, India. The remedial after school intervention comes from Banerjee et al. (2010), which increased letter recognition in Jaunpur district, India. The evidence from NGO-supported tracking programs is mixed: full-day tracking increased student learning in Western Province, Kenya (Duflo et al. 2011); partial day tracking did not increase learning in Bihar and Uttarakhand states, India (Banerjee et al. 2017); and partial day tracking increased learning when an extra supervisory layer and instructional hour accompanied it in Haryana state, India (Banerjee et al. 2017).

the design of the teaching, learning, and training materials and continued through training and implementation, creating a truly government owned and operated program. The increase in test scores demonstrates the potential potency of the interventions if implemented elsewhere entirely within a government system.Yet, we also show that continuing support beyond program inception is also crucial—the assistants' adherence fell over time.

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.

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. 2022). The official curriculum to which teachers must adhere and pedagogical methods that teachers use are largely unchanged from a time in which only wealthier, more highly educated parents could afford to send their children to school 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. This results in heterogeneous classrooms with many students left behind—only about a quarter of primary school students reach proficiency levels in English and math (Ministry of Education 2014). In our baseline data, 94 percent of grade 3 students could not read a grade 3 text, 18 percent of grade 3 students could not identify letters of the English alphabet, and the within grade-by-school heterogeneity was larger than the difference in the average test scores between grades 1 and 3.

In the year prior to the study, the language of instruction in lower primary grades changed from each school's discretion, usually a combination of English and a local language, to the school's assigned National Literacy Acceleration Program (NALAP) language.[3] Full implementation of this policy lingered into our study years (Hartwell 2010). Because of the NALAP delays, our analysis focuses on math and English skills, providing separate estimates for NALAP test scores.

2.2 National Youth Employment Program

The National Youth Employment Program (NYEP) paid the intervention's assistants, known as Teacher Community Assistants (TCAs). 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 NYEP. In preparation for the implementation, Ghanaian education officials visited India to learn from Pratham, a large Indian NGO, about the previous successes and challenges of the Teaching at the Right Level (TaRL) approach that was studied in Banerjee et al. 2007,

[3] A school's NALAP language was determined by geography and was not necessarily the mother tongue of all or a majority of the schools' students.

2010, and 2017. Government employees under the Ministry of Education umbrella designed the teaching, learning, and training materials with inspiration from the TaRL approach.

This study tested four methods of improving student learning in government schools—pull-out remedial, after school remedial, classroom split, and partial day tracking—relative to each other and a control group. 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]

Each intervention involved an educator, i.e., the person who teaches the pedagogy to the students. Schools in the three assistant-based treatments—pull-out remedial, after school remedial, and classroom split—used the same hiring procedures to hire an assistant who would be paid through NYEP. School Management Committees (SMCs) and Parent Teacher Associations (PTAs) identified potential assistants from secondary school graduates aged 18 to 35 living in the school community. Candidates were interviewed and selected for employment by a panel of local, GES, and NYEP representatives. In the partial day tracking intervention, the educators were existing classroom teachers in grades 1 through 3.

Existing government trainers provided all educators the same training on how to engage in active, child-focused pedagogy and materials that contained suggested engaging, child-focused activities.[4] Assistants in the remedial arms received additional training materials for remedial instruction. Classroom split assistants adhered to the official, grade level, curriculum. Teachers in the partial day tracking intervention received materials that spanned remedial to grade level to allow them to differentiate instruction across three learning levels. All educators were responsible for their own lesson plans with the provided materials as suggestions and guides.

[4] In active pedagogy, children take an active role in their own learning instead of passively receiving knowledge.

Educators were to implement the program at each primary school for one hour each day, four days per week. Educators received training on how to divide the students appropriately depending on the intervention. In the remedial and partial day tracking interventions educators tested students at the start of each term to determine their learning levels, assigning each student to learning level 1, 2 or 3. Remedial assistants worked remedial students across the three lower primary grades. This resulted in smaller, more homogeneous classrooms for all students for part of each day in the pull-out remedial intervention. Remedial students in the after school remedial intervention received an extra instructional hour. Assistants in the classroom split worked with a random half of the students from a classroom on grade level material. They were to randomly pick students each day. This provided all students a smaller class size. The partial day tracking teachers divided students by learning level within their classrooms in the first two terms of the intervention. Starting the third term of implementation, teachers learned through a refresher training to divide their students across grades by learning level with one teacher teaching each learning level.[5] Students in this intervention had a more homogeneous classroom environment. 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.

All interventions had the same timing and implementation schedule and occurred over three academic years. Initial trainings occurred in May (Term 3) of the 2010-2011 academic year (academic year 1) 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 (academic year 3).

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.

[5] For example, the grade 1 teacher might work with level 1 students, the grade 2 teacher with level 2 students, and the grade 3 teachers with level 3 students. In both the remedial interventions and the partial day tracking, learners were grouped with peers at their learning level but not necessarily their grade level.

[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 received the intervention starting in the third term grade 2, was treated for all of grade 3, and we tested 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, more homogeneous class versus an extra instructional hour to deliver remedial material. Both of these interventions were designed to shift the left tail of the learning distribution to the right with the pull-out version also providing a smaller class size and more homogeneous learning environment to all learners. 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 pull-out remedial and the partial day tracking interventions (T1 vs T4) shows whether a classroom teacher can replicate the benefits of a smaller, homogeneous class size by focusing on a homogeneous group of learners. When comparing the after school remedial to the classroom 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 after school remedial relative to the partial day tracking (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 classroom split relative to the partial day tracking (T3 vs T4) shows the relative effect of a smaller class size versus a more homogeneous learning environment. These last two interventions were designed to shift the entire test score distribution to the right, not only focusing on remedial learners.

4 Empirical Strategy

From our randomization design, comparing outcomes between individuals in treatment and control schools is straightforward. We estimate an overall effect size across the four treatments in an intent-to-treat specification,

$$y_{is} = \alpha + \beta \, treatment_s + X_{is}^{'}\Gamma + \varepsilon_{is} \tag{1}$$

where y_{is} is outcome y for individual i in school s, $treatment_s$ is an indicator variable equal to one if school s was a treatment school with a single indicator for all 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. 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}. 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.[6]

We additionally estimate the effect of each treatment separately,

$$y_{is} = \alpha + \sum_{T=1}^{4} \beta_T \, treatment_{Ts} + X_{is}^{'}\Gamma + \varepsilon_{is} \tag{2}$$

[6] Our point estimates are similar in magnitude but less precisely measured if we omit the baseline test scores as a covariate.

with separate indicators $treatment_{Ts}$ for each treatment T (the control group is the omitted category) and other notation as above.

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; on teachers' and assistants' attendance, time on task, and material usage; and on the likelihood the groups were meeting as intended.

Because of imperfect fidelity of implementation, we also perform an instrumental variables analysis of the treatment on the treated (TOT). In this case, assignment to treatment at the school level is the instrument for whether we observed correctly formed groups during the spot check sessions. We then follow the analog to the above specifications, first estimating the overall effect of the treatments then estimating the effects separately.

5 Sample Selection and Data

The 500 school experimental sample was nationwide in scope, including schools from all ten regions and 42 districts in Ghana.[7] From this sample, one hundred schools were randomly allocated into each of the five treatment designations (four treatment arms and a control group), stratified by region, above/below median average baseline student test score, and above/below median pupil teacher ratio.

To evaluate the effect of the four interventions, we collected nine rounds of data across three academic years: 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. The baseline occurred near the start of academic year 1 (October 2010), the first achievement

[7] In Ghana, district is the administrative subdivision immediately below region. Forty-two (out of 170 at the time) districts were randomly selected with at least two districts selected from each of the ten regions. The number of districts was limited to facilitate training educators from multiple schools at the same time, as would happen in a nationwide scale-up of the program. Each of the 42 districts was randomly assigned to have either 11 or 12 sampled schools. Within each district, sample schools were selected from Ghana's Education Management Information Systems (EMIS) school list, attempting to have an equal number of urban and rural schools.

follow-up was in academic year 2 (November 2011), and the second achievement follow-up was near the end of academic year 3 (July 2013). In academic year 1, we randomly sampled 25 students from grades 1 and 2 from those present on the day of initial enumeration. We attempted to follow these students through academic year 3 when they should have been in grades 3 and 4 if they progressed on pace.[8] The six spot-check rounds occurred termly, starting with the third term of academic year 1 (June 2010) and ending with the second term of academic year 3 (April 2013). In these data collection rounds, we visited a sub-sample of schools and recorded whether the school was implementing the intended intervention, assistant demographics, classroom activities, whether the student was still attending the particular school and in the expected grade, and student, teacher and head teacher attendance. Figure 2 above shows the data collection timeline. 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). To provide some context, almost all students had shoes but only one quarter had a clean, good quality uniform. Even though the official age of entry for grade 1 is 6, grade 1 students were on average 7.8 years old near the start of the academic year. 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 were about 36 years old and about half were women. On average each grade had 37 students and one teacher. About 30 percent of schools had electricity. Assistants were about 25 years old and 40 percent were women. Almost all had completed high school and about one third aspired to teach in the future.

Baseline achievement levels were low and heterogeneous within schools. At baseline, only about one half of grade 1 students could correctly name a presented English letter and one third could perform simple 1-digit addition. At baseline, the average standard deviation

[8] Students were encouraged to come to school on the days of the achievement follow-ups. Enumerators attempted to follow-up with all students who were absent, even those who had moved or were attending another school. In the first achievement follow-up an additional cohort of grade 1 students was added to the data collection. As these students could have selectively matriculated to school based on treatment status, we only use their data to calculate the between grade test score differences and not the treatment effect.

within grade 1 in a school was almost 90% of the average score difference between grades 1 and 2.

6 Results

6.1 Student Outcomes

Achievement, Selection into the Test, and Persistence

Table 1 contains the effects of the four treatments on the combined math and English student test scores, i.e. Equation 1 with a student's test score as the outcome. The sample is students who were grade 1 in academic year 1 when the treatment started and should have been in grade 2 in academic year 2 and grade 3 in academic year 3.[9] Panel A combines all interventions into a single treatment indicator. Panel B contains separate estimates for each of the four interventions. The first two columns are from the academic year 2 follow-up, after only about two terms of treatment. Relative to the control group, the treatment schools increased test scores by 0.08 SD (Panel A, column 1). Most of this gain was the result of the two remedial interventions. The pull-out remedial and after school remedial interventions increased test scores by a statistically significant 0.11SD (Panel B, column 1). The other two interventions, i.e. classroom split and partial day tracking, increased test scores by a positive, but statistically insignificant, 0.05 and 0.06 SD. We fail to reject that all the interventions had the same test score effect. At this follow-up students in the control schools in grade 2 had test scores that were 0.66SD higher than grade 1 students. Therefore, the overall effect across all interventions of 0.08SD is about 12 percent of a grade level of learning. As part of the motivation of the interventions was to improve foundational literacy and numeracy, in column 2 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 that has been used to evaluate similar interventions in India.[10] The interventions on average

[9] We attempted to interview and assess all baseline students regardless of their grade at follow-up.

[10] The ASER uses four types of questions to assess a student's reading level: reading letters, words,

increased test scores by 0.11SD. When only considering these foundational questions, the test score increases between the separate interventions are statistically significantly different: two remedial interventions have larger, and statistically different, point values than the the classroom split.

[Table 1 about here]

In academic year 3, students should have been in grade 3 and have received the intervention for two full school years. Columns 3 and 4 contain estimates of the effect of the interventions on academic year 3 achievement. When considered together, the interventions increased learning 0.11SD (Panel A, column 3). As with academic year 2, the largest point values are for after school remedial (0.15SD) and pull-out remedial (0.14SD) (Panel B, column 3). Unlike in academic year 2, both the classroom split and partial day tracking increased achievement by a statistically significant amount (0.08SD). When limiting the analysis to the foundational questions, the point values are larger (0.13SD on average, Panel A, column 4), and we reject the equality of the remedial interventions relative to the classroom split (0.15SD remedial vs. 0.07SD for the classroom split, Panel B, column 4).[11] Part of the theory of change with these remedial lessons is that they provide a strong foundation on which to build grade-level material knowledge. While the point values are larger for these remedial interventions for the grade level knowledge, they are only statistically different from the classroom split when limited to the foundational skills. At this follow-up, students in control schools who were in grade 3 had test scores that were 0.42SD higher than students in grade 2. Therefore, the effect size of 0.11SD is about 27 percent of a grade level of learning, over twice the grade level adjusted learning from academic year 2.

One concern with any randomized controlled trial is that the effects are artificially generated by differential selection into test taking based on treatment status. We found no

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.

[11] Appendix Table A4 contains subject-specific test score effects for both the entire test (Panel A) and the foundational content only (Panel B).

differential selection by treatment status or the interaction of treatment status and baseline achievement (Appendix Table A5). Nevertheless, we provide Lee (2009) bounds in Appendix Table A6, finding similar results as those in Table 1.

In Table 2, we test for the persistent effects of the intervention on students who should have received the program in grades 2 and 3 and been in grade 4 at the academic year 3 follow-up, and thus one year removed from the program. When considering all questions, i.e. grade levels 1-4, the pooled effect size is 0.07SD (column 1) with specific effect sizes ranging from (statistically insignificant) 0.01SD (partial day tracking) to (statistically significant) 0.12SD (classroom split). We reject that the classroom split and partial day tracking are statistically equal. Therefore, while the average effect persisted, the tracking intervention no longer increased test scores as much as the classroom split. The focus in the classroom split on grade-level content could have prepared students better for grade 4 content than the tracking or remedial interventions. The next column limits the exam to questions from the grade 1 through 3 curriculum. Receiving any treatment increased scores on average 0.10SD with point estimates for the specific treatment from (statistically insignificant) 0.05SD (partial day tracking) to (statistically significant) 0.13SD (classroom split) (column 2). We reject that the classroom split and partial day tracking had the same effect at the 10 percent level. The persistent effects on foundational material (column 3) are similar to the foundational gains for students one year removed from the *balsakhi* program in Banerjee et al. (2007) even though the immediate effects after two years of these interventions were about half the size of the *balsakhi* program.

[Table 2 about here]

Non-cognitive Outcomes

Based on the data from our unannounced spot-checks, we test for the effect of the intervention on three non-cognitive outcomes: absenteeism, no longer attending school, and grade repetition (Jackson 2018). For each student we calculate the average portion of days absent across all spot checks, whether the school ever reported that they were no longer attending that school (the sum of dropping out and transferring), or whether at any spot check they were in a grade below what would be expected based on timely progression. In control schools, students were absent about 36 percent of days, 24 percent were ever reported as no longer attending that school, and of those who were still attending the school, 23 percent were in a grade below their expected grade. The interventions did not change these non-cognitive outcomes. The full point values appear in Appendix Table A7.

Heterogeneity by Baseline Characteristics

The analysis thus far focused on the test scores of all students. Both partial day tracking and classroom split changed instruction for all students. Non-remedial students could have benefited from the remedial interventions through more homogeneous learning times during the pull-out remedial lessons or from having classmates closer to grade-level due to the after school remedial lessons. The heterogeneity analyses in Appendix Table A8 and Appendix Figure A1 confirms this uniformity of effect. We test for heterogeneity by baseline test score and whether the learners with in the top two-thirds of their grade, an approximation for not needing remedial attention. Whether considering all interventions together or each intervention separately, the coefficients on the interaction between baseline student achievement and treatments are statistically insignificant (Table A8, Panels A and B, columns 1 and 2). Appendix Figure A1 shows the non-parametric distributional effects of the interventions combined (sub-figure a) and then each intervention separately relative to the control group (sub-figures b through e). The effects of the interventions were mostly positive for students across the baseline test score distribution. One reason why we might find homogeneous effects across the baseline score distribution is that even the top students had limited literacy and numeracy at baseline. At the baseline, 54 percent of grade 1 students could read an upper or lower case English letter, 7 percent could read a three letter word, 76 percent could recognize a one digit number, and 37 percent could do one digit addition.

The intervention was not designed to favor one gender, yet gender might be a salient concern in a country with a gender bias in the assessment of teachers by head teachers (Beg, Fitzpatrick, and Lucas 2021). Based on a simple comparison of means for the analysis sample, girls' test scores were 0.06SD lower than boys' scores at baseline. At the year 3 follow-up, this difference had widened to 0.10SD for the control group, yet in the treatment groups, girls' test scores were 0.03SD higher than boys' scores. In column 3 of Table A8 we formally test for heterogeneity in effects on the academic year 3 test scores by student gender by interacting the treatment variables with female. When using a single treatment indicator, the effect on boys' test scores is 0.07SD with girls' test scores increasing by an additional 0.10SD (Table A8, Panel A, column 3). For the three interventions with a remedial or tracking component, girls' test scores increased statistically more than boys' test scores by about 0.10SD for the pull-out and after school remedial and 0.15SD for partial day tracking. Girls appear to have benefited more from homogeneous classrooms than boys, perhaps because girls were more hesitant to speak up to ask or answer questions in a heterogeneous learning environment. As teachers were 10 percentage points more likely than the assistants to be women, the additional improvement for girls in the assistant-led arms is not likely due to gender-matching role model effects.

6.2 Implementation and Treatment on the Treated

The estimates in Section 6.1 were intention to treat (ITT) estimates. Not all schools implemented the groups as intended or implemented them consistently, likely scenarios for other government implemented programs. In this sub-section, we use our spot-check data to measure the extent to which implementation occurred and then use assignment to treatment as an instrument for a school implementing the program in a two stage least squares estimation.

The emphasis of the interventions was grouping students, whether by remedial status, by learning level, or to have a smaller class size. In column 1 of Table 3 we test schools' fidelity of implementation of group learning over the spot-check rounds. Each spot-check overlapped with the time in which group learning should have been occurring in treatment schools. For each school, the portion of visits that the school was observed implementing group learning is the dependent variable. Overall, being in a treatment group increased the likelihood of students being divided into groups by 27 percentage points (Table 3, Panel A, column 1). The level of implementation was statistically different for each intervention, ranging from 6 percent of the time (partial day tracking) to 41 percent (after school remedial), with the other two interventions in between (pull-out remedial at 34 percent and classroom split at 27 percent). We did not observe any control schools grouping their students.

[Table 3 about here]

Since groups did not meet as frequently as prescribed, students did not get the dosage intended. In the remainder of Table 3 we implement a two stage least squares strategy with treatment status at the school level as an instrument for receiving the treatment, defined as the portion of spot-check visits in which students were correctly grouped. As with the previous results, we first combine all interventions into a single treatment indicator. These IV results should be interpreted with caution as they rely on two assumptions that might not hold for all interventions. First, they assume linearity and scale in dosage, e.g. a 10 percentage point increase in group learning is the same regardless of the base and doubling the portion of times that group learning was observed should double the effect sizes. Second, the exclusion restriction requires that the interventions only affect test scores through group learning. We show below that these assumptions likely hold for the three assistant interventions.

Based on the instrumental variables (IV) estimates, the combined interventions increased student test scores 0.30SD across all questions in academic year 2 (Panel A, column 2) and 0.41SD across all questions in academic year 3 (column 4). As with the non-instrumented version, the point values are larger when considering the foundational questions (0.40SD in academic year 2 and 0.46SD in academic year 3, columns 3 and 5). Each of these estimates is about 3.6 times the size of the non-IV estimates in Table 1 and range from 45 to 122 percent of a year of learning. Panel B contains the coefficient estimates for each intervention individually using the four school level treatment status assignments as instruments. The two remedial interventions are the only two interventions with statistically significant score increases for the entire test in academic year 2 (0.31SD and 0.27SD, column 1). The remedial interventions also increased test scores in academic year 2 for the foundational content by about 0.36SD (column 2). Based on this IV strategy, the partial day tracking increased foundational test scores by 1.8SD (column 2), and the increase is statistically different than the other three interventions (see the discussion below for some of the reasons why the partial day tracking coefficient might be a biased estimate of the true effect). In academic year 3, for the entire test, the pull-out remedial (0.42SD), after school remedial (0.36SD), and classroom split (0.29SD) all increased overall test scores (column 4). The remedial interventions also increased foundational test scores in year 3 (0.43SD and 0.39SD) as did the partial day tracking (2.1SD).

The IV estimates of the effects of the assistant led interventions are robust to specifications designed to test the assumptions about linearity and scale in dosage and the exclusion restriction. First, to address concerns about linearity and scale in dosage, in Appendix Table A9 we redefine whether group learning was occurring as an indicator variable that takes the value of 1 if schools were observed implementing group learning in at least 50 percent of observations and 0 otherwise. Among the treatment schools, about 29 percent of schools were observed grouping their students at least half the time. When using this binary measure of implementation, the partial day tracking intervention no longer has a statistically significant effect on achievement, nor can we reject that its effects are equal to the other interventions. The other IV results are robust to this modification. Second, to address concerns about the exclusion restriction, we first show in Appendix Table A10 that the interventions at most minimally changed the operations of schools and classrooms outside of group learning, with the exception of those schools assigned to partial day tracking.[12] Therefore, the exclusion restriction is likely satisfied for the assistant interventions but less likely to hold for the partial day tracking. Appendix Table A11 provides revised IV estimates removing all schools

[12] In schools assigned to partial day tracking, the head teacher was more likely to be present and teachers were more likely to be in the classroom, engaged with students, and using materials. Therefore, students in this intervention likely received more effective teaching even when group learning was not happening.

assigned to the partial day tracking intervention. The IV results from Table 3 are robust to this change in sample. One final concern could be measurement error in the portion of time that group learning occurred as we only observed schools occasionally and not daily. Any measurement error should be unrelated to treatment status as the protocol for visiting each school was the same.

7 Cost Effectiveness

We base the cost effectiveness on the costs of the program using the ingredients method as the program was designed.[13] Based on estimates of scaling a single program to the entire country and assuming similarly sized schools to the study, the per student annual costs would be $19.60 for the remedial assistant interventions, $18.77 for the classroom split, and $10.65 for the partial day tracking intervention. 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 classroom split, 0.36SD for the partial day tracking, and 0.38SD for the pull-out or after school remedial. The similarity of the cost effectiveness of the three remedial or tracking interventions are remarkable—the assistant-led ones cost approximately twice as much per student with approximately twice the benefit.[14]

Because we have a multiple year intervention considering the cost effectiveness of a shorter duration of the program is tempting. The first follow-up occurred near the start of the second school year, about two terms into the program. The effect size for the overall score at this first follow-up was about 71 percent of the point values for the second follow-up and the costs were less than half. Based on this metric, a shorter duration of the program could be

[13] This could overstate the realized costs of the program as some expenses scheduled for year 1 occurred later in the program.

[14] As is common in the literature, this evaluates each intervention at the point value of its estimated effect size.

considered more cost effective. If instead one considers the percent of a year of learning, then the point value in academic year 3 is 27 percent of a year of learning, over twice the years of learning of 12 percent in academic year 2.

8 Discussion

8.1 Implementation Lessons

Because of our extensive data collection over three academic years, this study contributes to the understanding of potential pitfalls, challenges, and successes when implementing something entirely within existing government systems. Weaker oversight, delayed materials, and inconsistent payments were three ways in which having this program implemented by the government likely contributed to smaller effect sizes than could have been possible.

Common Challenges and Successes Across All Intervention

Common challenges likely muted effects across all four interventions. Any differences between levels of implementation or student test score effects across interventions cannot be the result of these common attributes. Almost all educators were trained as intended. Refresher trainings occurred throughout the study for all educators. Material delivery was delayed equally across the interventions. The program used existing education sector production and distribution systems for material delivery. At the start of academic year 2, only 12 percent of schools had received materials. All educators were subject to the same existing school environments with weak oversight and school leaders who might have been hesitant or skeptical about the merits of the interventions and had a strong focus on completing the annual curriculum.

Unique Challenges and Successes For Each Intervention

Because of the design of each intervention, other challenges and successes were intervention-specific, differentially affecting the level of implementation and resulting student test score changes.

First, how were the teacher and the assistant interventions different? Teachers' had a dual mandate of teaching learning levels during partial day tracking and teaching grade level material in regular curriculum lessons while assistants only had the singular mandate to teach their group lessons. When teachers should have been engaging in partial day tracking, we most often observed them teaching curriculum lessons to their regular classrooms. Teachers' employment through GES created weaker incentives than the assistants' annual contracts. Teachers were, however, more likely to be paid on time. Assistant salaries were delayed across all interventions, especially in years 2 and 3. Taken together, teachers were the most likely educator to be present but were the least likely to implement. The likelihood that the assistant interventions met as prescribed decreased over time while the likelihood that the teacher-led partial day tracking occurred increased (Appendix Table A12).

Second, how were the assistant interventions from each other? The more similar the intervention was to status quo teaching operations, the more likely the assistant was absorbed into those operations instead of performing group teaching as intended. Classroom split assistants covered for a classroom teacher about 20 percent of the time they were present. This happened about half as often to the pull-out remedial assistants and almost never happened to the after school remedial assistants. Therefore, the classroom split intervention occurred less frequently than the other assistant interventions even though the classroom split assistants were more likely to be present.

8.2 Mechanisms

In this section we discuss the mechansisms behind improvements in student learning using the conceptual framework and comparisons acros the interventions from Section 3.2 and Figure 1. Unfortunately, based on this study the relative effect sizes of perfect implementation cannot be known, instead the realized implementation is what could be expected when implemented within existing government systems.

For simplicity we focus on the findings from academic year 3. The remedial interventions had the largest and most similar point values (0.14SD and 0.15SD). Once remedial instruction was included, a smaller class size had almost the same point value as an extra instructional hour (0.008SD difference) potentially because the status quo in-class work was at an inappropriate learning level for students who needed remedial instruction. By design, the two remedial interventions focused primarily on learners at the lower end of the baseline achievement distribution while the other two interventions sought to increase test scores throughout the distribution. The two remedial interventions had statistically larger effects than the classroom split on foundational content—working with fewer learners on grade level content is less effective at increasing foundational learning that remedial focused instruction. This foundational content was something many learners needed, not only those at the bottom of the baseline score distribution. One appeal of partial day tracking is that it teaches all learning levels, yet its average effects were statistically indistinguishable and smaller than for the remedial interventions. Recall that the partial day tracking intervention was implemented about one fifth as often as the other interventions. Had it been implemented to the same degree as the other interventions, the effect sizes might have exceeded the others'. The instrumental variables estimates that scale the effects relative to the actual dose received have larger coefficients, averaging 0.4SD, yet as with the intention to treat estimates, the effect sizes are statistically indistinguishable from each other on the full exam.

8.3 Comparison to Other Studies

Recall from Section 1 that this study combined many of the lessons of the last 20 years in improving student learning into a single RCT implemented within existing government systems. Based on the previous studies implemented within an existing government system (Bold et al. 2018) or without an extra supervisory layer (Banerjee et al. 2017), whether any of the four interventions when implemented by the government would increase student learning was unclear. In Appendix Table A13 we compare some of the common steps along the causal chain that are reported across studies—educators being trained, being present, using intervention materials, and adhering to the intervention as measured by whether the correct learners were being taught with the prescribed method—between the previous implementations of class size reductions, full and partial day tracking, and pull out and after school remedial instruction.[15] The educators in this study were more likely to be trained but less likely to be present. Relative to previous implementations that improved test scores, they were less likely to be using intervention materials or adhering to the interventions.

Appendix Figure A2 plots our effect sizes (solid bars) relative to previous interventions that are the most similar to ours. 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. We show that existing systems can increase learning.

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 natiowide RCT in Ghana 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 randomlu 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

[15] Blanks in the table denote that the study did not report this statistic while "N/A" indicates that the particular study did not include this element. For example, the tracking and classroom split interventions in Kenya did not include materials.

of grade-level for a part of each day.

Showing that governments can improve productivity in government primary schools, all four interventions increased student learning based on test administered at the end of grade 3 for those students who started the program near the end of grade 1. The average effect of the treatments was 0.11SD. The interventions' positive effects persisted for those students exposed to the program grades 2 and 3 and tested at the end of grade 4, one year after ending the program. We find no evidence that the program affected the non-cognitive outcomes of student attendance, drop-out, or likelihood of being demoted. Taking into account imperfect compliance by using a treatment on the treated (TOT) estimate, the intervention increased test scores an average of 0.4SD. When considering cost effectiveness from the intention to treat estimates, the after school remedial, pull-out remedial, and partial day tracking interventions were similarly cost effective—the effect sizes and costs of the first two were approximately twice the size of the third.

All interventions 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 intervention as prescribed 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.

Bau, N. and J. Das (forthcoming). Teacher value-added in a low-income country. *American Economic Journal: Policy*.

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 (forthcoming). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Policy*.

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 P. Niehaus (2017). Experimentation at scale. *Journal of Economic Perspectives 31*(4), 103–24.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2016). Building state capacity: Evidence from biometric smartcards in india. *American Economic Review 106*(10), 2895–2929.

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.[16] 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 were 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. Students with oral test scores are included in the analysis. Within round achievement is converted to latent scores using item response theory and standardized using the control group mean

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

and standard deviation for the grade 1 baseline cohort 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 across the interventions with the after school remedial assistants most likely to live in the community.

Students

[Appendix Table A1 about here]

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 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. Panel A includes all test questions while Panel B focuses on the questions most similar to the ASER test in India. Ghana introduced a new local language program the year prior to the intervention whose full implementation occurred during the intervention years (columns 5 and 6).

[Table A4 about here]

In Table A5 we test for differential selection into test taking by treatment status, finding none.

[Table A5 about here]

Despite finding no evidence of differential selection into test taking by treatment status, Table A6 provides Lee (2009) bounds for our estimates.

[Table A6 about here]

Table A7 shows that the intervention had no effect on students' non-cognitive outcomes—being absent on an unannounced day, no longer attending the baseline school, or repeating a grade.

[Table A7 about here]

Table A8 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 A8 about here]

Tables A9 through A11 provides additional IV estimates and related robustness checks.

[Table A9 about here]

[Table A10 about here]

[Table A11 about here]

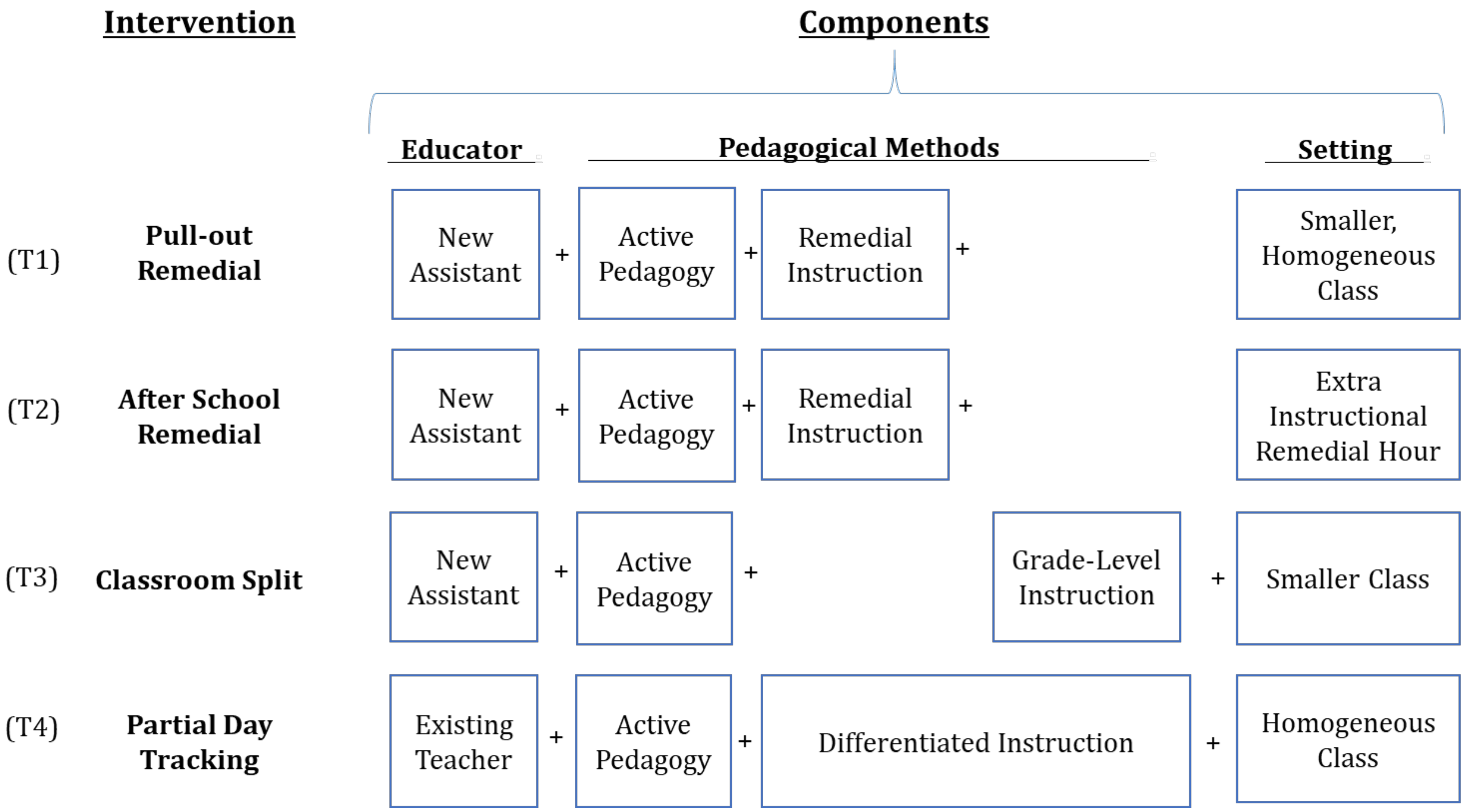
Table A12 shows the evolution of implementation over time.

[Table A12 about here]

Table A13 provides the effect sizes from from this paper, rows denoted (1), and the other studies in Figure A2 along with the steps along the causal chain–educators being trained, being present, using the intervention materials, and adhering to the intervention. Jim Berry graciously provided the combined math and literacy test scores for the Banerjee et al. (2016) set of interventions.

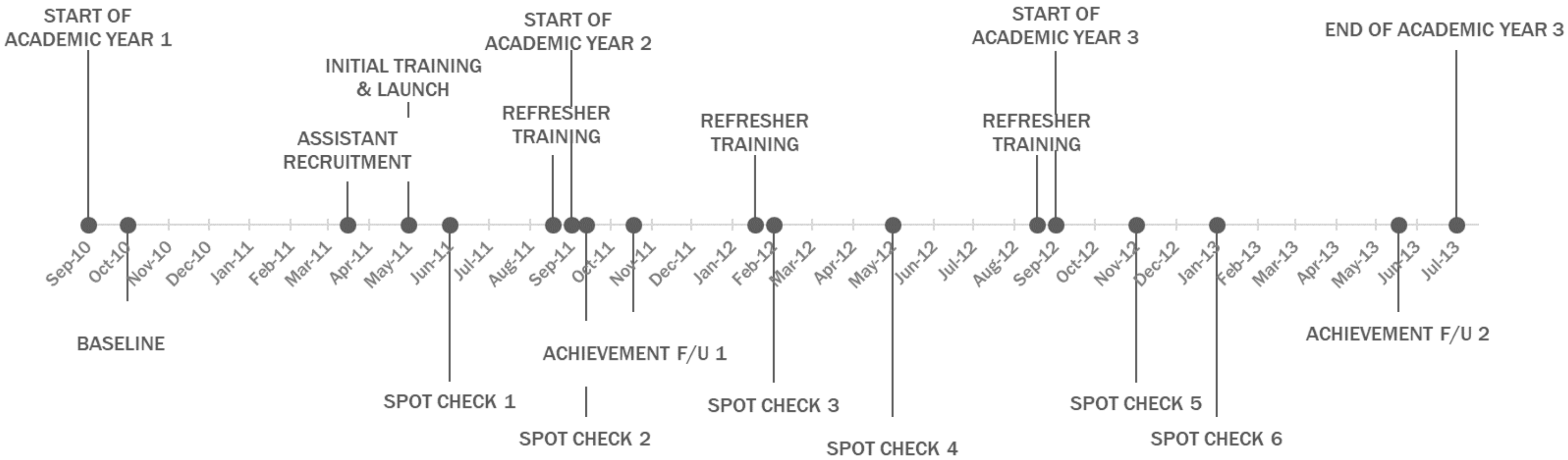
[Table A13 about here]

Figure 1: Intervention Components and Graphical Conceptual Framework



	Intervention	Components				
		Educator	Pedagogical Methods			Setting
(T1)	**Pull-out Remedial**	New Assistant +	Active Pedagogy +	Remedial Instruction +		Smaller, Homogeneous Class
(T2)	**After School Remedial**	New Assistant +	Active Pedagogy +	Remedial Instruction +		Extra Instructional Remedial Hour
(T3)	**Classroom Split**	New Assistant +	Active Pedagogy +		Grade-Level Instruction +	Smaller Class
(T4)	**Partial Day Tracking**	Existing Teacher +	Active Pedagogy +	Differentiated Instruction +		Homogeneous Class

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.

Table 1: Effects on Achievement in Math and English

	Academic Year 2		Academic Year 3	
	All Questions	Foundational Questions	All Questions	Foundational Questions
	(1)	(2)	(3)	(4)
Panel A: Interventions Combined				
Any Intervention	0.080**	0.113***	0.113***	0.125***
	(0.034)	(0.032)	(0.035)	(0.035)
Observations	8,654	8,654	8,004	8,004
R-squared	0.54	0.46	0.47	0.39
Panel B: Interventions Separately				
(1) Pull-out Remedial	0.106**	0.137***	0.143***	0.147***
	(0.043)	(0.040)	(0.047)	(0.048)
(2) After School Remedial	0.110**	0.143***	0.151***	0.154***
	(0.046)	(0.043)	(0.046)	(0.046)
(3) Classroom Split	0.047	0.066*	0.082*	0.068
	(0.045)	(0.040)	(0.044)	(0.043)
(4) Partial Day Tracking	0.059	0.110**	0.077*	0.135***
	(0.046)	(0.043)	(0.046)	(0.044)
P-value of Test of Equality				
1 = 2	0.93	0.90	0.88	0.90
1 = 3	0.20	0.08	0.20	0.10
1 = 4	0.32	0.54	0.18	0.80
2 = 3	0.20	0.08	0.15	0.07
2 = 4	0.30	0.48	0.13	0.69
3 = 4	0.80	0.31	0.92	0.14
Observations	8,654	8,654	8,004	8,004
R-squared	0.55	0.46	0.47	0.39
Test score difference between grades	0.66	0.57	0.42	0.38

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 in grade 1 tested at baseline (academic year 1) and relevant follow-up round. Difference between grades calcuated based on control group means between grades 1 and 2 (AY2) and grades 2 and 3 (AY3). Columns 2 and 4: questions most similar to those appearing in the ASER. See text for more details.

Table 2: Persistent Achievement Effects in Math and English

	All Questions	Grade 1-3 Questions	Foundational Questions
	(1)	(2)	(3)
Panel A: Interventions Combined			
Any Intervention	0.074**	0.102***	0.104***
	(0.036)	(0.034)	(0.035)
Observations	4,302	4,302	4,302
R-squared	0.49	0.43	0.42
Panel B: Interventions Separately			
(1) Pull-out Remedial	0.072	0.102**	0.110**
	(0.045)	(0.043)	(0.044)
(2) After School Remedial	0.086*	0.114***	0.118***
	(0.046)	(0.044)	(0.045)
(3) Classroom Split	0.120**	0.134***	0.133***
	(0.049)	(0.046)	(0.048)
(4) Partial Day Tracking	0.014	0.052	0.053
	(0.049)	(0.046)	(0.047)
P-value of Test of Equality			
1 = 2	0.77	0.79	0.86
1 = 3	0.33	0.49	0.64
1 = 4	0.22	0.27	0.23
2 = 3	0.50	0.68	0.76
2 = 4	0.15	0.19	0.18
3 = 4	0.04	0.10	0.12
Observations	4,302	4,302	4,302
R-squared	0.49	0.43	0.42
Test score difference between grades	0.45	0.47	0.51

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: Grade 2 students tested at baseline (academic year 1) and in academic year 3. Students progressing on pace stopped receiving the intervention at the end of academic year 2. Test was invigilated at the end of academic year 3.

Table 3: Fidelity of Implementation and Instrumental Variable Effects

	Target Educator Teaching to a Group	Instrumental Variable Estimates			
		Academic Year 2		Academic Year 3	
		All Questions	Foundational Questions	All Questions	Foundational Questions
	(1)	(2)	(3)	(4)	(5)
Panel A: Interventions Combined					
Any Intervention	0.268***	0.296**	0.401***	0.411***	0.463***
	(0.017)	(0.127)	(0.121)	(0.127)	(0.127)
Observations	500	8,654	8,654	8,004	8,004
R-squared	0.23	0.54	0.45	0.47	0.38
Panel B: Interventions Separately					
(1) Pull-out Remedial	0.340***	0.307**	0.376***	0.417***	0.431***
	(0.027)	(0.124)	(0.122)	(0.135)	(0.138)
(2) After School Remedial	0.410***	0.266**	0.363***	0.360***	0.387***
	(0.033)	(0.110)	(0.104)	(0.112)	(0.114)
(3) Classroom Split	0.270***	0.168	0.206	0.288*	0.253
	(0.028)	(0.164)	(0.149)	(0.155)	(0.154)
(4) Partial Day Tracking	0.055***	1.064	1.774**	1.273	2.078**
	(0.015)	(0.829)	(0.849)	(0.784)	(0.819)
P-value of Test of Equality					
1 = 2	0.09	0.74	0.92	0.67	0.75
1 = 3	0.06	0.37	0.23	0.40	0.25
1 = 4	0.00	0.33	0.08	0.25	0.03
2 = 3	0.00	0.52	0.25	0.62	0.35
2 = 4	0.00	0.31	0.08	0.22	0.03
3 = 4	0.00	0.25	0.05	0.18	0.02
Observations	500	8,654	8,654	8,654	8,654
R-squared	0.40	0.54	0.54	0.54	0.54
Control Group Mean or Test Score Difference	0.00	0.66	0.57	0.42	0.38

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Additional controls: strata. Column 1: at the school level. Dependent variable defined as the portion of spot-checks in which the target educator was teaching to the intended group. Columns 2-5: Additional controls: baseline test score and female. Instrumental variable estimates with treatment assignment at the school level as an instrument for groups meeting. See caveats in the text regarding the estimates for Partial Day Tracking.

Appendix Table A1: Summary Statistics--Students

	Treatment				Control	Test of Equality F-stat (p-value)
	Remedial Instruction		Classroom Split	Partial Day Tracking		
	Pull-out	After School				
	(1)	(2)	(3)	(4)	(5)	(6)
Combined English and Math Test Score	-0.038 (0.93)	-0.008 (0.99)	0.029 (1.01)	-0.051 (0.92)	0.000 (1.00)	0.24 (0.92)
English Test Score	-0.030 (0.96)	-0.011 (1.02)	0.031 (1.04)	-0.078 (0.94)	0.000 (1.00)	0.41 (0.8)
Math Test Score	-0.037 (0.93)	-0.005 (0.97)	0.022 (0.99)	-0.015 (0.93)	0.000 (1.00)	0.13 (0.97)
Local Language Test Score	-0.046 (1.00)	-0.051 (0.99)	0.014 (0.93)	-0.039 (0.95)	0.000 (1.00)	0.20 (0.94)
Student wore a clean, good quality uniform	0.25 (0.44)	0.27 (0.45)	0.28 (0.45)	0.21 (0.41)	0.24 (0.43)	1.77 (0.13)
Student had shoes	0.88 (0.32)	0.90 (0.30)	0.89 (0.31)	0.88 (0.32)	0.89 (0.31)	0.16 (0.96)
Student Age	7.68 (0.32)	7.64 (0.30)	7.85 (0.31)	7.86 (0.32)	7.82 (0.31)	0.00 (0.96)
Father Literate	0.47 (0.50)	0.47 (0.50)	0.46 (0.50)	0.43 (0.49)	0.44 (0.50)	0.68 (0.61)
Mother Literate	0.32 (0.47)	0.35 (0.48)	0.35 (0.48)	0.32 (0.47)	0.31 (0.46)	0.50 (0.74)
Someone at home helps with homework	0.60 (0.49)	0.58 (0.49)	0.63 (0.48)	0.59 (0.49)	0.61 (0.49)	0.86 (0.40)
Self-reported absences in the last week	0.76 (1.35)	0.79 (1.38)	0.87 (1.44)	0.86 (1.46)	0.83 (1.40)	0.64 (0.63)

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 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					Test of Equality
	Remedial Instruction		Classroom Split	Partial Day Tracking	Control	F-stat (p-value)
	Pull-out	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 teachers who completed a baseline survey. Panel B: Averages calculated over study schools.

Appendix Table A3: Summary Statistics--Assistant Characteristics

	Treatment			Test of Equality F-stat (p-value)
	Remedial Instruction		Classroom Split	
	Pull-out	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 NALAP Language	0.91 (0.29)	0.92 (0.28)	0.91 (0.29)	0.08 (0.93)
Speak Most Common Student Primary Language	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 Effects

	English		Math		Local Language	
	Academic Year 2	Academic Year 3	Academic Year 2	Academic Year 3	Academic Year 2	Academic Year 3
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Including Grade-Level Content						
Panel A1: Interventions Combined						
Any Intervention	0.063*	0.122***	0.095***	0.094***	0.083*	0.077*
	(0.036)	(0.040)	(0.034)	(0.033)	(0.044)	(0.045)
Observations	8,654	8,004	8,654	8,004	8,654	8,002
R-squared	0.49	0.45	0.46	0.36	0.40	0.39
Panel A2: Interventions Separately						
(1) Pull-out Remedial	0.090*	0.141***	0.115***	0.133***	0.113**	0.118**
	(0.047)	(0.051)	(0.041)	(0.045)	(0.056)	(0.057)
(2) After School Remedial	0.099**	0.178***	0.111**	0.108**	0.074	0.076
	(0.049)	(0.051)	(0.043)	(0.044)	(0.060)	(0.060)
(3) Classroom Split	0.036	0.075	0.066	0.058	0.061	0.053
	(0.045)	(0.048)	(0.042)	(0.041)	(0.056)	(0.057)
(4) Partial Day Tracking	0.029	0.097**	0.091**	0.076*	0.084	0.060
	(0.049)	(0.049)	(0.044)	(0.042)	(0.057)	(0.061)
P-value of Test of Equality						
1 = 2	0.85	0.47	0.92	0.61	0.52	0.50
1 = 3	0.26	0.19	0.21	0.11	0.35	0.26
1 = 4	0.22	0.38	0.55	0.23	0.61	0.36
2 = 3	0.20	0.04	0.29	0.27	0.83	0.70
2 = 4	0.18	0.10	0.64	0.49	0.88	0.80
3 = 4	0.88	0.64	0.56	0.67	0.69	0.91
Observations	8,654	8,004	8,654	8,004	8,654	8,002
R-squared	0.49	0.45	0.46	0.36	0.40	0.39
Panel B: Foundational Content Only						
Panel B1: Interventions Combined						
Any Intervention	0.107***	0.131***	0.086***	0.099***	0.130***	0.137***
	(0.033)	(0.037)	(0.032)	(0.032)	(0.042)	(0.042)
Observations	8,654	8,004	8,654	8,004	8,654	8,002
R-squared	0.33	0.33	0.41	0.32	0.23	0.28
Panel B2: Interventions Separately						
(1) Pull-out Remedial	0.119***	0.130**	0.112***	0.141***	0.134**	0.168***
	(0.045)	(0.051)	(0.040)	(0.044)	(0.055)	(0.056)
(2) After School Remedial	0.166***	0.183***	0.100**	0.109**	0.184***	0.176***
	(0.045)	(0.049)	(0.042)	(0.043)	(0.057)	(0.056)
(3) Classroom Split	0.043	0.079*	0.060	0.053	0.104**	0.069
	(0.042)	(0.046)	(0.040)	(0.041)	(0.052)	(0.050)
(4) Partial Day Tracking	0.103**	0.134***	0.074*	0.095**	0.096*	0.138***
	(0.045)	(0.045)	(0.043)	(0.042)	(0.052)	(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: Selection Into Test Taking

	Test Taker in			
	Academic Year 2		Academic Year 3	
	(1)	(2)	(3)	(4)
Pull-out Remedial	0.024 (0.018)	0.027 (0.019)	0.014 (0.018)	0.004 (0.018)
Pull-out Remedial X BL Score		0.004 (0.023)		-0.017 (0.026)
After School Remedial	0.013 (0.017)	0.003 (0.017)	0.000 (0.017)	-0.016 (0.020)
After School Remedial X BL Score		-0.017 (0.022)		-0.030 (0.024)
Classroom Split	0.028 (0.017)	0.015 (0.018)	0.007 (0.017)	-0.007 (0.020)
Classroom Split X BL Score		-0.023 (0.022)		-0.026 (0.024)
Partial Day Tracking	-0.023 (0.021)	-0.021 (0.022)	-0.012 (0.018)	-0.029 (0.020)
Partial Day Tracking X BL Score		0.003 (0.029)		-0.030 (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. Linear probability models.

Appendix Table A6: Lee Bounds

	Combined English and Math		English		Math		Local Language	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Academic Year 2								
Pull-out Remedial	0.113***	0.118***	0.087*	0.093**	0.117***	0.121***	0.100*	0.114**
	(0.043)	(0.042)	(0.047)	(0.047)	(0.041)	(0.039)	(0.056)	(0.053)
After School Remedial	0.121***	0.120***	0.102**	0.103**	0.117***	0.114***	0.074	0.092
	(0.045)	(0.045)	(0.048)	(0.049)	(0.043)	(0.043)	(0.060)	(0.058)
Classroom Split	0.052	0.065	0.028	0.041	0.065	0.077*	0.070	0.068
	(0.043)	(0.043)	(0.045)	(0.045)	(0.042)	(0.041)	(0.056)	(0.054)
Partial Day Tracking	0.064	0.064	0.031	0.030	0.085**	0.086*	0.086	0.084
	(0.045)	(0.046)	(0.048)	(0.048)	(0.043)	(0.044)	(0.056)	(0.057)
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: Academic Year 3								
Pull-out Remedial	0.141***	0.157***	0.132**	0.146***	0.131***	0.146***	0.104*	0.120**
	(0.048)	(0.047)	(0.051)	(0.051)	(0.046)	(0.045)	(0.057)	(0.056)
After School Remedial	0.152***	0.155***	0.175***	0.177***	0.107**	0.112**	0.074	0.084
	(0.047)	(0.047)	(0.051)	(0.051)	(0.044)	(0.044)	(0.060)	(0.059)
Classroom Split	0.065	0.076*	0.067	0.079	0.054	0.062	0.051	0.055
	(0.044)	(0.043)	(0.048)	(0.048)	(0.041)	(0.040)	(0.056)	(0.055)
Partial Day Tracking	0.088**	0.093**	0.092*	0.097**	0.072*	0.076*	0.057	0.063
	(0.044)	(0.045)	(0.049)	(0.049)	(0.042)	(0.042)	(0.061)	(0.061)
Observations	7,854	7,853	7,854	7,853	7,854	7,853	7,852	7,851
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 A7: Non-cognitive Outcomes

		Student Was	
	Absent	No Longer Attending	Repeating a Grade
	(1)	(2)	(3)
Panel A: Interventions Combined			
Any Intervention	0.002	0.018	0.021
	(0.016)	(0.014)	(0.025)
Observations	10,952	10,803	9,446
R-squared	0.04	0.03	0.04
Panel B: Interventions Separately			
(1) Pull-out Remedial	0.001	0.005	-0.002
	(0.023)	(0.018)	(0.030)
(2) After School Remedial	0.004	0.023	0.030
	(0.020)	(0.018)	(0.030)
(3) Classroom Split	-0.008	0.017	0.014
	(0.020)	(0.017)	(0.029)
(4) Partial Day Tracking	0.011	0.028	0.043
	(0.020)	(0.019)	(0.031)
P-value of Test of Equality			
1 = 2	0.88	0.33	0.26
1 = 3	0.70	0.50	0.58
1 = 4	0.66	0.24	0.13
2 = 3	0.54	0.73	0.55
2 = 4	0.75	0.78	0.66
3 = 4	0.36	0.55	0.31
Observations	10,952	10,803	9,446
R-squared	0.04	0.03	0.04
Control Group Mean	0.36	0.24	0.23

Notes : * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school appear in parenthesis. Additional controls: strata and female. Sample: grade 1 students at baseline. Column 1: Asked in all spot check rounds. Portion of days absent across all rounds. Column 2 and 3: Asked in spot check rounds 2-6. Equals 1 if the school ever reported outcome at the top of the column. Linear probability models. Column 2: For some absent students in some rounds the respondents did not know if a student planned to return--those rounds for those students are not included in this estimation. Column 3: Responses only recorded for students still attending the school in that round.

Appendix Table A8: Heterogeneity in Achievement Effects

		By Baseline Achievement		
		Test Score	Top Two Thirds of School by Grade	By Gender
		(1)	(2)	(3)
Panel A: Interventions Combined				
	Any Intervention	0.116*** (0.035)	0.085* (0.046)	0.066* (0.039)
	Any Intervention X BL Achievement	-0.000 (0.034)	0.040 (0.046)	
	Any Intervention X Female			0.099** (0.045)
	BL Achievement or Female (varies by column)	0.492*** (0.031)	-0.007 (0.044)	-0.021 (0.041)
	Observations	8,004	8,004	8,004
	R-squared	0.46	0.47	0.47
Panel B: Interventions Separately				
(1)	Pull-out Remedial	0.143*** (0.046)	0.108 (0.066)	0.090* (0.052)
(2)	Pull-out Remedial X BL Ability	0.033 (0.043)	0.050 (0.064)	
(2)	Pull-out Remedial X Female			0.109** (0.055)
(3)	After School Remedial	0.153*** (0.046)	0.105* (0.060)	0.103* (0.053)
(4)	After School Remedial X BL Ability	-0.016 (0.046)	0.064 (0.059)	
(4)	After School Remedial X Female			0.100* (0.057)
(5)	Classroom Split	0.082* (0.044)	0.036 (0.058)	0.063 (0.051)
(6)	Classroom Split X BL Ability	0.021 (0.043)	0.065 (0.056)	
(6)	Classroom Split X Female			0.039 (0.057)
(7)	Partial Day Tracking	0.084* (0.046)	0.093 (0.059)	0.006 (0.051)
(8)	Partial Day Tracking X BL Ability	-0.042 (0.046)	-0.023 (0.061)	
(8)	Partial Day Tracking X Female			0.149** (0.060)
	BL Ability or Female (varies by column)	0.491*** (0.032)	-0.008 (0.044)	-0.021 (0.041)
	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		0.17	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 A9: Additional Instrumental Variables Estimates: Redefining Fidelity of Implementation

	Teaching to a Group at Least Half of Observations	Instrumental Variable Estimates			
		Academic Year 2		Academic Year 3	
		All Questions	Foundational Questions	All Questions	Foundational Questions
	(1)	(2)	(3)	(4)	(5)
Panel A: Interventions Combined					
Any Intervention	0.285***	0.276**	0.374***	0.390***	0.439***
	(0.026)	(0.120)	(0.116)	(0.122)	(0.122)
Observations	500	8,654	8,654	8,004	8,004
R-squared	0.17	0.54	0.44	0.45	0.37
Panel B: Interventions Separately					
(1) Pull-out Remedial	0.384***	0.261**	0.317***	0.347***	0.351***
	(0.050)	(0.115)	(0.121)	(0.124)	(0.133)
(2) After School Remedial	0.456***	0.238**	0.325***	0.321***	0.343***
	(0.051)	(0.102)	(0.101)	(0.106)	(0.113)
(3) Classroom Split	0.279***	0.172	0.215	0.300*	0.272
	(0.045)	(0.168)	(0.164)	(0.171)	(0.181)
(4) Partial Day Tracking	0.023	2.080	3.455	2.801	4.543
	(0.022)	(1.938)	(2.372)	(2.173)	(2.776)
P-value of Test of Equality					
1 = 2	0.30	0.84	0.95	0.83	0.96
1 = 3	0.10	0.57	0.50	0.77	0.65
1 = 4	0.00	0.34	0.18	0.25	0.13
2 = 3	0.01	0.67	0.45	0.89	0.66
2 = 4	0.00	0.33	0.18	0.25	0.13
3 = 4	0.00	0.31	0.16	0.24	0.12
Observations	500	8,654	8,654	8,654	8,654
R-squared	0.29	0.52	0.52	0.52	0.52
Control Group Mean	0.00	0.66	0.57	0.42	0.38

Notes : * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Additional controls: strata and round. Column 2: additional control: portion of teachers present at baseline at this school. 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.

Appendix Table A10: School and Classroom Operational Changes

	Head Teacher Present	Teachers			
		Present	In Classroom	Engaged with Students	Using Materials
	(1)	(2)	(3)	(4)	(5)
Panel A: Interventions Combined					
Any Intervention	0.075**	0.007	0.035	0.057**	0.084***
	(0.031)	(0.019)	(0.023)	(0.023)	(0.024)
Observations	1,892	6,324	6,305	6,143	2,378
R-squared	0.06	0.06	0.08	0.07	0.11
Panel B: Interventions Separately					
(1) Pull-out Remedial	0.135***	0.015	0.041	0.050*	0.025
	(0.038)	(0.025)	(0.029)	(0.029)	(0.028)
(2) After School Remedial	0.058	0.003	0.024	0.017	0.007
	(0.039)	(0.025)	(0.028)	(0.026)	(0.029)
(3) Classroom Split	0.018	-0.029	0.015	0.047*	0.018
	(0.039)	(0.027)	(0.029)	(0.028)	(0.028)
(4) Partial Day Tracking	0.089**	0.037	0.057**	0.110***	0.242***
	(0.039)	(0.024)	(0.029)	(0.028)	(0.032)
P-value of Test of Equality					
1 = 2	0.05	0.66	0.55	0.21	0.51
1 = 3	0.00	0.11	0.37	0.91	0.78
1 = 4	0.24	0.36	0.58	0.04	0.00
2 = 3	0.30	0.24	0.75	0.22	0.70
2 = 4	0.44	0.18	0.24	0.00	0.00
3 = 4	0.07	0.01	0.14	0.02	0.00
Observations	1,892	6,324	6,305	6,143	2,378
R-squared	0.07	0.06	0.08	0.07	0.16
Control Group Mean	0.51	0.69	0.52	0.36	0.05

Notes : * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Additional controls: strata and round. Column 2: additional control: portion of teachers present at baseline at this school. Columns 3 and 4: not conditional on teacher being present. Column 5: conditional on a teacher being present.

Appendix Table A11: Additional Instrumental Variables Estimates: Limited Sample

	Target Educator Teaching to a Group	Instrumental Variable Estimates			
		Academic Year 2		Academic Year 3	
		All Questions	Foundational Questions	All Questions	Foundational Questions
	(1)	(2)	(3)	(4)	(5)
Panel A: Interventions Combined, Omitting Partial Day Tracking					
Group Learning	0.338***	0.250**	0.323***	0.351***	0.362***
	(0.019)	(0.105)	(0.099)	(0.103)	(0.103)
Observations	400	6,993	6,993	6,440	6,440
R-squared	0.35	0.57	0.48	0.48	0.40
Panel B: Interventions Separately, Omitting Partial Day Tracking					
(1) Pull-out Remedial Groups	0.337***	0.303**	0.376***	0.400***	0.425***
	(0.028)	(0.128)	(0.123)	(0.135)	(0.137)
(2) After School Remedial Groups	0.411***	0.261**	0.359***	0.357***	0.391***
	(0.033)	(0.109)	(0.102)	(0.110)	(0.111)
(3) Classroom Split Groups	0.269***	0.170	0.209	0.287*	0.251*
	(0.029)	(0.164)	(0.148)	(0.151)	(0.149)
P-value of Test of Equality					
1 = 2	0.07	0.74	0.89	0.75	0.80
1 = 3	0.07	0.38	0.23	0.46	0.25
2 = 3	0.00	0.54	0.26	0.62	0.33
Observations	400	6,993	6,993	6,440	6,440
R-squared	0.37	0.57	0.48	0.48	0.40
Test score difference between grades		0.66	0.57	0.42	0.38

Notes : * significant at 10%; ** significant at 5%; *** significant at 1%. Outcomes are standard deviation test score changes. School level treatment status as an instrument for portion of times that groups occurred. 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 in grade 1 tested at baseline (academic year 1) and relevant follow-up round. Difference between grades calcuated based on control group means between grades 1 and 2 (AY2) and grades 2 and 3 (AY3).

Appendix Table A12: Unannounced Visits--Fidelity of Implementation Over Time

		Target Educator Teaching to a Group	
	Academic Year 1	Academic Year 2	Academic Year 3
	(1)	(2)	(3)
Panel A: Interventions Combined			
Any Intervention	0.360*** (0.030)	0.271*** (0.022)	0.172*** (0.022)
Observations	450	494	472
R-squared	0.17	0.18	0.11
Panel B: Interventions Separately			
(1) Pull-out Remedial	0.488*** (0.053)	0.322*** (0.041)	0.198*** (0.040)
(2) After School Remedial	0.527*** (0.053)	0.413*** (0.044)	0.310*** (0.048)
(3) Classroom Split	0.357*** (0.052)	0.285*** (0.037)	0.112*** (0.034)
(4) Partial Day Tracking	0.040 (0.033)	0.066*** (0.025)	0.059** (0.028)
P-value of Test of Equality			
1 = 2	0.59	0.12	0.07
1 = 3	0.06	0.49	0.09
1 = 4	0.00	0.00	0.00
2 = 3	0.02	0.02	0.00
2 = 4	0.00	0.00	0.00
3 = 4	0.00	0.00	0.21
Observations	450	494	472
R-squared	0.31	0.28	0.17
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 A13: Other Studies--Effect Sizes and Compliance

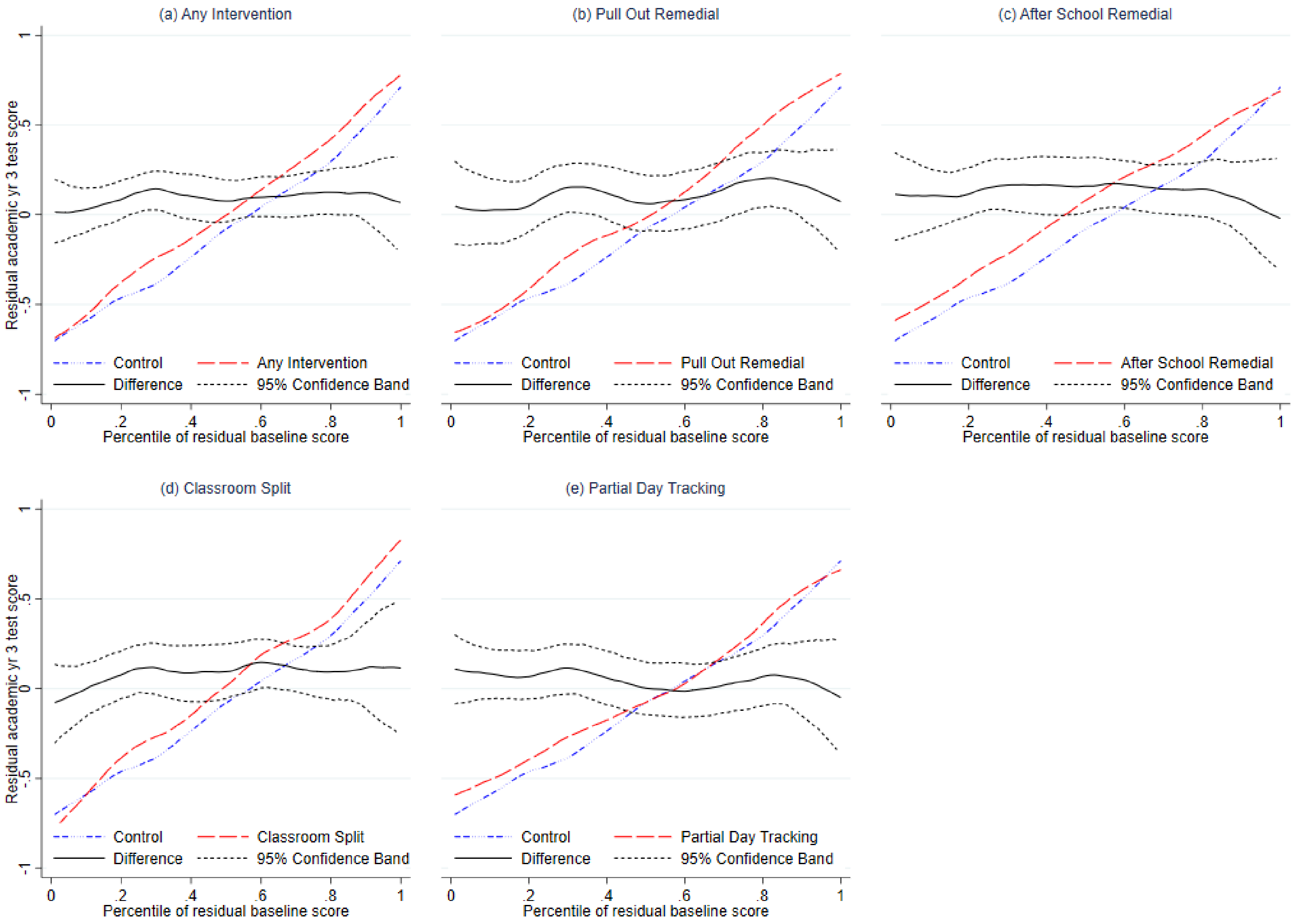
Intervention	Location	Effect on Student Test Scores	Educator				Duration (years)
			Trained	Present	Using Intervention Materials	Groups Meeting	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: This Study							
(1) All Interventions Combined	Ghana	0.113*** (0.035)	0.95	0.59	0.15	0.26	2
(1) Pull-out Remedial	Ghana	0.143*** (0.047)	0.99	0.58	0.38	0.33	2
(1) After School Remedial	Ghana	0.151*** (0.046)	0.98	0.48	0.34	0.41	2
(1) Classroom Split	Ghana	0.082* (0.044)	0.97	0.63	0.39	0.27	2
(1) Partial Day Tracking	Ghana	0.077* (0.046)	0.85	0.73	0.11	0.06	2
Panel B: Other Government Implementation							
(2) Contract Teacher, Classroom Split	Kenya	0.021 (0.090)	N/A	0.73	N/A	0.71	1.5
Panel C: NGO Supported							
(3) Pull-out Remedial	Mumbai + Vadodara,	0.284** (0.060)					2
(4) After School Remedial+Village Education Committee Training	Jaunpur district, India	(a)					1
(5) Partial Day Tracking	Bihar state, India	0.0316 (0.0369)	0.67		0.58	0.04	1
(5) Partial Day Tracking	Uttarakhand state, India	0.0264 (0.0340)	0.28		0.26	0.10	1
(5) Partial Day Tracking+extra supervisory layer+devoted new school hour	Haryana state, India	0.0862*** (0.0161)	0.96		0.74	0.92	1
(5) Partial Day Tracking+Assistant	Uttarakhand state, India	-0.00195 (0.0315)	0.45		0.34	0.06	1
(5) Partial Day Tracking+After School Remedial	Bihar state, India	0.124*** (0.0344)	0.67		0.64		1
(6) Contract Teacher, Full Year Tracking	Western Province,	0.176** (0.077)	N/A	0.85	N/A	0.99	1.5
(7) Contract Teacher, Full Year Split	Western Province,	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.

Appendix Figure A1: Non-Parametric Distributional Effects



(a) Any Intervention

(b) Pull Out Remedial

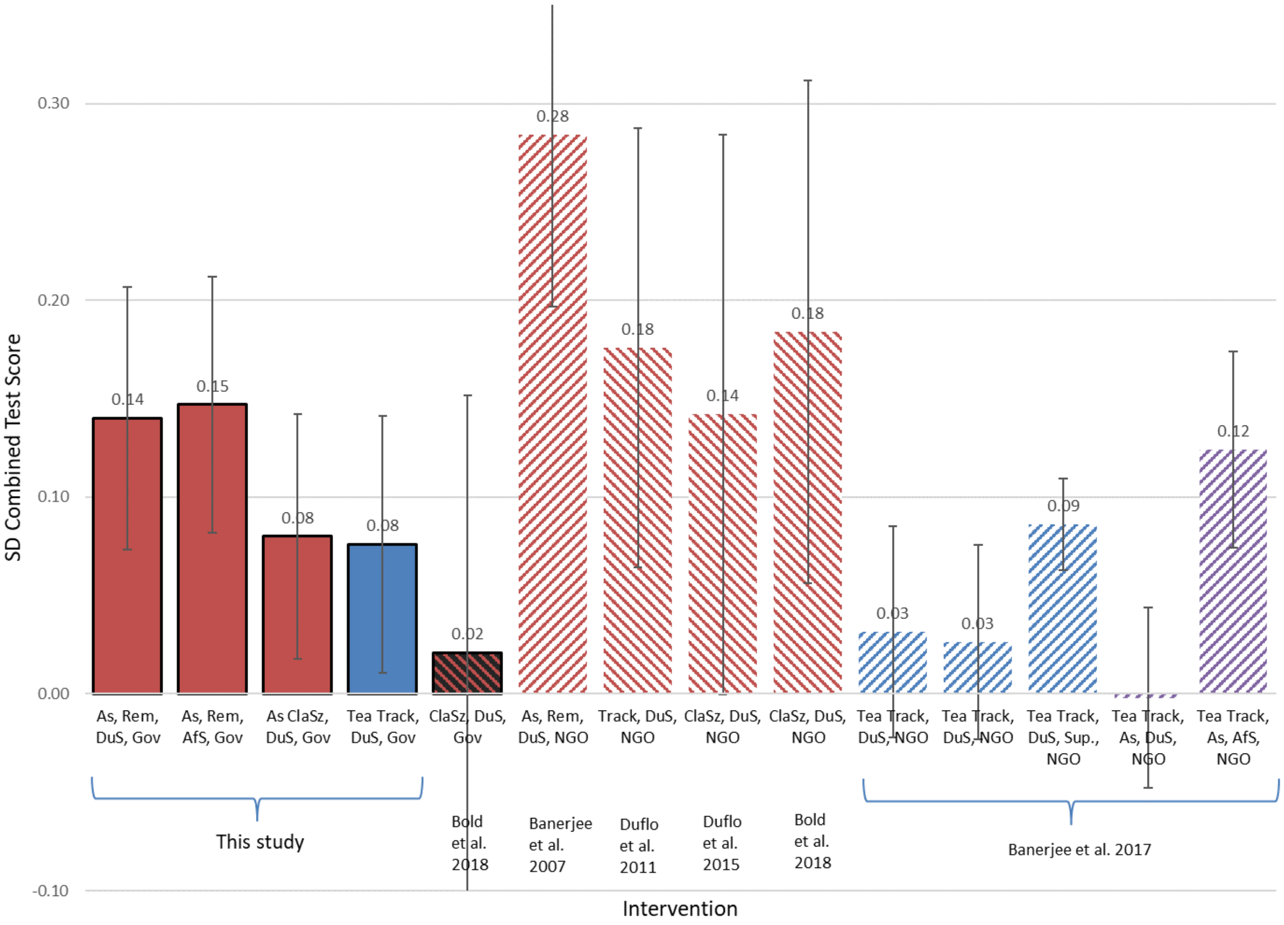
(c) After School Remedial

(d) Classroom Split

(e) Partial Day Tracking

Notes: Panel A: all interventions combined. Panels B-E: the specific treatment relative to the control group.

Figure A2: 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 existing 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. ClaSz=smaller class size. Track=partial or full day tracking. Sup=extra supervisory layer. Gov=government implemented. NGO=NGO implemented.