External Validity: Four Models of Improving Student Achievement

Annie Duflo, Jessica Kiessel, and Adrienne M. Lucas*

May 22, 2020

Abstract

Randomized controlled trials in lower-income countries have demonstrated ways to increase learning, in specific settings. This study uses a large-scale, nationwide RCT in Ghana to show the external validity of four school-based interventions inspired by other RCTs. Even though the government implemented the programs within existing systems, student learning increased across all four models, more so for female than male students, and many gains persisted one year after the program ended. Three of the four interventions had a similar cost effectiveness. The intervention that directly targeted classroom teachers increased the likelihood that teachers were engaged with students.

*This paper was previously circulated under the title, "Every Child Counts: Adapting and Evaluating Targeted Instruction Approaches into a New Context through a Nationwide Randomized Experiment in Ghana." Duflo: Innovations for Poverty Action. Kiessel: Omidyar Network. Lucas (corresponding author): Department of Economics, University of Delaware, 419 Purnell Hall, Newark, DE 19716, CGD, J-PAL, and NBER. alucas@udel.edu. 302.831.1901. We gratefully acknowledge generous funding for the evaluation from the International Growth Centre, the Hewlett Foundation, and the Children's Investment Fund Foundation. Many thanks to the Amma Aboagye, Albert Akoubila, and Maame Araba Nketsiah for supporting and championing the implementation of program and to Ama Anaman, Raphael Bandim, Suvojit Chattopadhyay, Callie Lowenstein, Sam N'tsua, and Pace Phillips for outstanding research implementation and project management. We would also like to thank Wendy Abt for her instrumental role in getting this project started, and Caitlin Tulloch and Shahana Hirji for their leadership and support with the cost analysis. For research assistance, we thank Joyce Jumpah, Ryan Knight, Harrison Diamond Pollock, and Matthew White. We also acknowledge our partners at the Ministry of Education, Ghana Education Services, and the Ministry of Youth Sports and Culture without whom this project would not have been possible. We thank David Evans and Fei Yuan for providing statistics on existing impact evaluations. For useful comments and suggestions, we thank Noam Angrist, Sabrin Beg, Jim Berry, Anne Fitzpatrick, John Floretta, Sarah Kabay, Heidi McAnnally-Linz, Jeremy Tobacman, and seminar participants at Swarthmore College, the University of Delaware, the Northeast Universities Development Consortium Conference, and the Teaching at the Right Level Conference. This RCT was registered in the American Economic Association Registry for randomized control trials as AEARCTR-0005912.

JEL Codes: I21, I25, I28, J24, O15.

Keywords: external validity, education, primary school, targeted instruction, teaching at the right level

1 Introduction

Even though many countries share the common challenge that students are enrolled in school but not learning (e.g. Andrabi et al. 2007; Muralidharan 2013; Jackson, Rockoff, and Staiger 2014; Bold et al. 2019) most rigorous studies to improve education focus on one or two interventions at a time, in a specific setting, often with a well resourced non-governmental implementing partner (Hanushek 2019; Evans and Yuan 2019).[1] Lessons are then broadly applied, in different settings with different systems, assuming external validity where potentially none exists. Even when government partners are eager to improve student outcomes based on evidence, the most effective way to achieve success within existing government accountability systems is often unclear. This study tests what happens when education interventions that were successful (and unsuccessful) elsewhere are brought to a new continent, implemented by government partners, and compared to each other. We implemented a 5-arm, 500 school randomized controlled trial across all 10 regions of Ghana to assess the external validity and effects, relative to each other and a control group, of four alternative instructional models to increase student achievement, all implemented within existing government structures. Similar models are currently being implemented in eleven African countries, and this is the first formal evaluation outside of India.

Specifically, we evaluated the Teacher Community Assistant Initiative (TCAI), a Ghana Ministry of Education program designed to test the relative efficacy of four alternative school-based interventions to increase student learning in grades 1 through 3, i.e. lower primary school, while working within the existing schooling and youth employment systems. The four interventions were 1) providing schools with assistants to work with remedial learners on a pull-out basis during the school day, 2) providing schools with assistants to work with remedial learners outside of the school day, 3) providing schools with assistants to work with half of the classroom each day on grade-level content, 4) having existing teachers focus

[1] Based on the 177 education impact evaluation studies published between 1980 and 2017 that appear in Evans and Yuan (2019) and calculations graciously provided by those authors, more than 50 percent of the studies had one treatment arm and the 95th percentile study had 3 treatment arms.

instruction on students' learning levels by dividing students into learning levels for part of each day. The fifth arm was a control group with no change to instructional delivery. From the 500 school sample, schools were randomized into one of four interventions or the control group. In addition to being of specific interest to the Ministry, the selection of these four interventions also provides insight into their external validity–each intervention had been previously evaluated in India, but not relative to each other, nor within existing government structures.

Each intervention used a different method to reach learners, many of whom were behind grade level, and began with training an educator in student-centered, active pedagogy and providing accompanying teaching and learning materials. Who received the training, when the teaching occurred, and the learning levels of the intended beneficiaries varied by intervention.[2]

In three of the interventions, schools hired assistants who were directly paid by the existing Ghana National Youth Employment Scheme. Some assistants had prior experience as teachers, but this was not an expectation nor requirement. In *assistant led remedial instruction during school*, the assistants worked with remedial learners for part of each school day on a pull-out basis. In addition to the basic training, these assistants received specific remedial focused training and materials. In *assistant led remedial instruction after school*, assistants were similarly trained but worked with remedial learners outside of the school day, mostly after school hours, both providing remedial instruction and effectively lengthening the school day for these learners. In the *assistant split* intervention, the assistant worked with an arbitrary half of a class for a part of each school day, focusing on grade-level content. These assistants could better focus on individual students in a smaller setting, but they did not specifically focus on remedial skills.

The *teacher-led targeted instruction* intervention trained existing classroom teachers to focus their teaching at the learning level of the students instead of the grade level for part

[2]All interventions started near the end of the 2010-2011 academic year and finished at the end of the 2012-2013 academic year.

of each school day.[3]

We find that all four interventions increased student achievement on tests that included content from all three grades of lower primary school. After being exposed to the program for 2 years, students who started the program in grade 1 increased their test scores by 0.08 to 0.15 standard deviations (SD), depending on the exact intervention, in all cases statistically significant.[4] These changes are equivalent to 18 to 34 percent of a year of schooling in this context. We we are unable to reject the statistical equivalent of the four interventions for the combined math and English score. For the English score alone, we find that the the assistant-led remedial after school intervention was statistically different than both the assistant split and the teacher-led interventions. When only considering foundational content, the test score increases were similar, but we reject that the assistant split is as effective as the assistant-led remedial after school.

Despite their different effect sizes, three of the four interventions were equally cost effective. The effect sizes of the two assistant led remedial interventions were about twice the size of the teacher-led targeted instruction intervention and would cost about twice as much at scale due to the extra costs of the assistant salaries. Therefore, when put on the common scale of effect per $100, their cost-effectiveness is approximately equivalent. The assistant split was the least cost effective.

The effects persisted for students who experienced the program for less time and were one year removed from the program. For students who were exposed for 1.3 years, starting in grade 2 and were one year removed from the program at follow-up, the effect sizes, including fourth grade content were 0.01 to 0.12 SD with similar effect sizes for foundational content. In the first year of the program, the only one these students experienced, interventions

[3]The exact method of dividing students by learning level changed during the course of the intervention. Teachers were first encouraged to divide their students within their classrooms by learning levels for part of the day. Teachers reported that keeping one group occupied while working with another group was difficult. Part way through the study, teachers were re-trained and instructed to divide their students by level across all three grades. Therefore, for part of each day, instead of instruction happening in grades 1 to 3, students were divided into learning levels 1 to 3.

[4]From largest to smallest, the point estimates were 0.15SD for assistant led remedial after-school, 0.14SD for assistant-led remedial during school, and 0.08SD for both assistant-split and teacher-led targeted instruction.

evolved and the materials were delayed, potentially explaining why for these older students the teacher-led model is no longer statistically significant.

The interventions were not designed to benefit one gender over another. Yet, when considering heterogeneity by gender, the test scores of female students increased by at least 0.10 SD more than for male students in the three interventions that had a remedial component, i.e. the two assistant-led remedial interventions and teacher-led targeted instruction.[5]

Beyond student learning, we collected data in 6 additional school visits. The program did not affect students' likelihood of being present, dropping out, or repeating a grade level. The intervention did not affect the likelihood that teachers were physically on campus, but teachers were 11 percentage points (31 percent) more likely to be engaged with students in the teacher-led targeted instruction arm. Conditional on being present, teachers in the teacher-led arm were twice as likely as those in the control group to have teaching and learning materials visible in their classrooms and 5 times as likely to be using teaching and learning materials. Therefore, even though the teacher-led model did not involve extra out of school time, students in the teacher-led model effectively received more instructional time.

These achievement and behavior changes occurred despite implementation difficulties: materials were initially delayed in reaching schools, assistant salaries were delayed throughout the program, and in the last spot check, 5 percent of schools were closed due to teacher strikes (unrelated to the program).

Our findings make four related contributions to the economics literature.

First, we contribute to a nascent literature on external validity and show that a commonality of challenges across contexts can trump differences between contexts when considering the applicability of an intervention. The question of external validity, how a program works outside the context of existing evaluations or even relative effectiveness within the same context, complicates decisions to scale up or even continue programs that appear promising.

Most education interventions are conducted in a particular setting, and rarely tested

[5]We find statistically significant increase in test scores across all interventions for girls (0.10 to 0.20 SD), but find statistically insignificant effects for boys subject to the assistant-split or teacher-led targeted instruction.

elsewhere despite an interest understanding which interventions can be scaled successfully across multiple settings (Deaton 2010; Meager 2019; Bo and Galiani 2019).[6] Social norms, institutions, and even the weather can mediate the effects of a program, causing those effects to vary by location, times, and the characteristics of beneficiaries. Context and program design are paramount in considering broad lessons from experimental impact evaluation (Pritchett and Sandefur 2015). Even geographically proximate countries can experience the same intervention differently (Lucas et al. 2014). Even though models of targeted instruction are currently being implemented in 11 African countries and reach over 5 million children in India, this is its first evaluation outside of India (Banerjee et al. 2017). India and Ghana share similar challenges in education—large heterogeneous classrooms with many learners behind grade level, yet the effects of programs could be quite different—especially because Ghana does not have a strong NGO to lead implementation as occurred in India. We show that context-specific complications might mitigate effect sizes, but challenges within the Ghanaian and Indian education sector were similar enough for programs to improve learning across both contexts. With improved fidelity of implementation, effect sizes in Ghana might even exceed those found in India.

Second, we test for validity across multiple interventions, simultaneously testing four models to improve student achievement relative to each other and a control group. Previous work simultaneously tested at most two related models simultaneously in the same location. Broadly in India, assistants who were hired and trained by Pratham, a large Indian NGO, improved student test scores, as did teachers when an additional Pratham created supervisory layer was added (Banerjee et al. 2007 ; Banerjee et al. 2016).[7] Also in India, teachers trained by Pratham but operating under their normal supervisory structure did not improve student test scores beyond the control group (Banerjee et al. 2010; Banerjee et al. 2016). In

[6]Two education studies that tested interventions across contexts are Lucas et al. (2014) that tested scaffolding pedagogy models in Kenya and Uganda and Bando et al. (2019) that tested inquiry- and problem-based pedagogy across four countries in Latin America.

[7]The results of Banerjee et al. (2016) were re-scaled and published in Banerjee et al. (2017). We refer to Banerjee et al. 2016 as it contains additional details on the interventions and the results appear in standard deviations.

Western Province, Kenya dividing students into smaller classes for the entire year did not statistically increase test scores on average (Duflo, Dupas, and Kremer 2015). Nor did it increase scores throughout Kenya when the Ministry of Education provided the extra teacher despite evidence on the success of similar programs in the United States (Krueger 1999; Bold et al. 2013). In contrast, for Kenyan students whose class size reduction was accompanied by the provision of a contract teacher hired by an NGO, test scores increased by about 0.2 SD (Bold et al. 2013; Duflo, Dupas, and Kremer 2015).[8] Dividing students by learning level for the entire school year in Western Province, Kenya also increased test scores (Dupas, Duflo, and Kremer 2012). In contrast to our study, these Kenyan interventions were for the entire school day and year and did not include teacher training. Appendix Table 1 contains additional details of these 5 studies in India and Kenya. In contrast with the existing results from the models most similar to the assistant split and teacher-led targeted instruction, we find positive effects of all four of our models tested.

Third, we contribute to the literature on working within schools to improve education in low income countries broadly. In addition to the specific targeted instruction papers cited above, other studies have shown the success, and sometimes complications, of programs to improve student learning by more efficiently using existing teaching resources through teacher training and accompanying materials or teacher incentives (e.g. scaffolding approaches to reading in Kenya and Uganda (Lucas et al. 2014); scripted lessons in Uganda (Kerwin and Thornton 2018); teacher incentives in Tanzania (Mbiti et al. 2019 and Mbiti, Romero, and Schipper 2019) and Uganda (Gilligan et al. 2018); see Muralidharan et al. 2017 for documentation of existing inefficiencies; see McEwan 2015 for a meta analysis on school-based interventions). The success of these programs that targeted teachers is in contrast to studies that have found that providing resources alone, e.g. textbooks or flip charts, did not increase student achievement, especially for those students who are not at the learning level to use these materials (e.g. Glewwe, Kremer, and Moulin 2004; Glewwe, Kremer, Moulin

[8]Muralidharan and Sundararaman (2013) find improved test scores with the addition of a contract teacher who could have been used to fill an existing teaching vacancy or reduce class size.

and Zitzewitz 2009; Banerjee et al. 2017). Further, another arm of the literature, primarily in South Asia and China, found that technology inside the classroom can improve learning (e.g. Banerjee et al. 2007; Beg et al. 2019).

Fourth, the intervention did not just work within schools, it worked within the existing schooling system without NGO support. Government ownership began at inception with a team traveling to India to learn from the experience of Pratham the large education NGO that implemented the targeted instruction and remedial programs studied in Banerjee et al. (2007), Banerjee et al. (2010), and Banerjee et al. (2017). Ghana Education Services and the Ministry of Education were the primary government agencies involved in designing the materials and training the teachers. Local schools hired the assistants who were paid through an existing youth employment scheme. Both the newly trained existing teachers and the new assistants were incorporated into schools with the existing, weak oversight system. The entire TCAI support apparatus outside of existing government personnel was a 4 person Technical Assistance Unit. The 400 treatment schools were equally divided into 4 geographic regions and one Regional Coordinator provided at most minimal support to 100 schools. The previous implementation of these interventions that increased test scores in India occurred along with heavy NGO involvement including NGO employees acting as trainers and recruiters and in some cases directly supervising teachers and teaching students. In Kenya, the NGO oversaw the hiring of the contract teachers and paid the salaries. Previous research portended disaster for these models when implemented by the government. Both Bold et al. (2013) and Kerwin and Thornton (2018) found that when programs that had been implemented by an NGO were implemented by the government, weak public institutions eliminated positive effects. Vivalt (2016) more broadly found programs that were implemented by academics or NGOs had larger effects that those that were government-implemented. One key difference in this study is that this did not start as an NGO initiative that was adopted (imperfectly) by the government, instead starting as a government program from the outset. While some of our point estimates are smaller than those previously

found, some are larger, demonstrating that government ownership of a program from the start can be a partial substitute for direct NGO supervision.

Taken together, we find strong external validity for the assistant-based models that previously increased learning and also encouraging findings for the teacher-led model that previously had no effect when implemented within existing government structures.

2 Background

We first present background on the schooling sector in Ghana broadly and then on the specific youth employment program that funded the assistants. This section closes by comparing the common problems between Ghana and India, the other country in which similar interventions have been tested.

2.1 The Ghanaian Educational System

Primary school in Ghana is grades 1 through 6, known as P1 through P6, and government primary schools should not charge fees.[9] Students should start primary school at age 6. The first three grades of primary school, P1-P3, are considered lower primary grades. Our study focuses on these students. The school year starts in September and consists of approximately three 13 week terms: mid-September through mid-December, early January through mid-April, and early May though the end of July.

In lower primary school, teachers are classroom teachers, teaching all subjects to the same group of students during a school year. The leader of the school is the head teacher, i.e. principal. Groups of 8 to 10 geographically proximate schools are overseen by a single Circuit Supervisor. Circuit Supervisors are government employees who are almost always former classroom teachers and report to the District Education Offices. Teachers' salaries

[9]After primary school, students continue on to junior high school for grades JHS1 to JHS3 (grades 7 through 9 in the US context). Secondary school is an additional three years (SHS1 through SHS3) and can be either vocational or academic track. During the period under study, government junior high school were fee-free but government secondary schools charged fees.

are paid centrally, and they are assigned to schools by the District Education Office.

The Ghanaian government spends approximately 30% of its budget on education, with nearly all of it going towards payroll. While primary school gross enrollment rates in Ghana were almost 98% in 2011, low achievement rates persisted like in many other countries in Africa. Even though the number of children in schools and the heterogeneity of their family background have increased substantially since the start of free primary education in 2005, the curriculum is largely unchanged from a time in which only wealthier, more highly educated parents could afford to send their children to school. Further, teachers are encouraged to adhere to the content and pace of the official curriculum regardless of the learning levels of their students. Many students in early primary lack the basic skills required to succeed in later school. Only about a quarter of students reach proficiency levels in English and Math, and approximately 42% fail to reach minimum competency in math or English (National Education Assessment 2013). Further, the proficiency levels vary by region with students in Greater Accra 4 times more likely to be proficient than students in the Northern region. Proficiency also varies within schools and heterogeneity increases as students reach higher grades.

A new language policy started in the school year immediately preceding our study, complicating our assessment of language learning. The National Literacy Acceleration Program (NALAP) assigned each school an official NALAP language based on its location. This language was not necessarily the most common language spoken by the students.[10] The primary language of instruction in lower primary grades, our sample grades, was to be the NALAP language with English introduced gradually in grade 1. By the end of grade 3 students were expected to be fluent in English and the NALAP language. The official language of instruction switches to English starting in grade 4. The NALAP implementation did not go

[10]Only 11 languages were considered NALAP languages. Students were expected to learn in the NALAP language even in heterogeneous areas or when a non-NALAP language was the most common language. In determining the NALAP language, students' actual use of language in the school was not considered. In our sample, for 37 percent of our sample schools, the most common language spoken by students was not the NALAP language. The NALAP language was one of the two most common languages for 82 percent of schools.

as planned with difficulties lingering into our study years with only 61 percent of our sample reporting having received NALAP materials and 75 percent of our teachers trained at our baseline, one year after the start of the NALAP program.[11]

Our study worked in schools with 9 of the 11 NALAP languages, and we developed appropriate testing tools in each language. Because of difficulties in the NALAP roll-out that lingered into our study years, our analysis focuses on math and English skills, providing estimates for the NALAP language separately.

2.2 National Youth Employment Program

The National Youth Employment Program (NYEP) paid the school-based assistants, known in the study as teacher community assistants (TCAs). The NYEP was an existing program under the Ministry of Youth and Sports that offered unemployed youth (18 to 35 years old), mostly secondary school graduates, two year public service positions and a small ($80-100) monthly stipend.[12] NYEP youth were already being used by the Ghana Education Service to fill vacant teacher positions, often in remote areas.[13]

2.3 Commonality of Challenges

As with many countries, both India and Ghana are in learning crises–students attend schools but are not learning. Nevertheless, teachers are encouraged to promote students to the next grade level and focus on grade-level curriculum, leading to heterogeneous classroom where students with the lowest learning levels are left behind (Gilligan et al. 2018). A series of studies in India have proposed various solutions to this crisis, testing one or two models

[11]Further complications included grade-level materials that assumed students had started learning their NALAP language in the first term of grade 1, yet no students had this foundation, and teachers lacking fluency in the NALAP language themselves (Hartwell 2010).

[12]NYEP was renamed the Ghana Youth Employment and Entrepreneurship Development Agency (GYEEDA) in 2012.

[13]This program is different than the Ministry of Education's National Service Scheme (NSS). The NSS is one year of national service, mandatory for all tertiary graduates. NSS participants are placed in both the private and public sectors, including schools. In our baseline, 12 percent of teachers were employed by NYEP and less than one percent by the NSS.

at a time, usually only successful with heavy involvement by Pratham, a large Indian non-governmental organization (NGO) (Banerjee et al. 2007; Banerjee et al. 2010; Banerjee et al. 2017). While many countries have similar issues to India of large heterogeneous classrooms where students are behind grade level, India is unique in having a large educational NGO to monitor, create, and support interventions to increase student learning. A typical low-income context has substantially lower quality oversight.

Ghana is a particularly salient location to test these interventions. Average within classroom heterogeneity is over two times the average achievement gap between grade levels, among grade 3 students only 28 percent reached the official proficiency level in English and 22 percent in math, and teachers are pressured to teach at the official level of the curriculum to prepare students for high stakes exams that occur at the end of junior high school (National Education Assessment 2013).

3 Intervention

In this section we first discuss each intervention arm in detail and then the logistics of the materials, personnel, and training.

3.1 Description of Treatments

This study tested four models of improving student learning relative to each other and a control group. Treatment was assigned at the school level with 100 schools receiving each treatment. Table 1 summarizes the components of each intervention. The interventions were not strictly nested but did contain common elements across multiple interventions.

[Table 1 about here]

Treatment 1: Assistant-led remedial instruction during school. In this pull-out program, assistants removed students from their regular classrooms to work on level-appropriate material. At the start of each term assistants tested students using a simple

tool to determine which students required remedial instruction and appropriately assign them a remedial level. Initially, assistants worked with multiple levels separately each day, e.g. Level 1 in the morning and Level 2 in the afternoon. In response to assistant feedback and monitoring visits, the model changed in early 2012 at the start of the second term of the second academic year. After this switch, assistants focused on one level at a time, i.e. bringing Level 1 students to Level 2 and then working with all Level 2 students. Assistants were expected to work with students 4 hours per week. Assistants received teaching and learning materials and training and were assigned a teacher within their school who was to act as their mentor.

Treatment 2: Assistant-led remedial instruction after school. This intervention was identical to treatment 1, but occurred outside of school hours, usually after school, for the same expected 4 hours per week. A shift similar to that in treatment 1 occurred in this treatment during academic year 2.

Treatment 3: Assistant-led random split during school. This intervention had the same number of contact hours as Treatments 1 and 2, but students were divided randomly instead of by learning level. Assistants observed the classroom teacher on Monday, when teachers introduced new material. For the rest of the week, the assistant worked with half of the class to review, reinforce, and teach the material, alternating which half. These assistants received materials and training that focused on student-centered learning, but not remedial instruction. They similarly had an assigned mentor teacher.

Treatment 4: Teacher led targeted instruction. Teachers used the same method as treatments 1 and 2 to determine each student's learning level. Also, as with treatments 1 and 2, the implementation changed during academic year 2. Initially, teachers were to divide the students within the classroom into three groups by level and address the needs of each group in a small group setting. At the start of 2012, the model changed and teachers combined their classes across grade levels and then divided them by learning level for one hour each day, four days per week, with each teacher covering a separate learning level and

focusing directly on the needs of that learning level.

Our primary cohort of interest was subject to these interventions (or part of the control group) starting with the third term of grade 1. They continued with these interventions through the end of grade 3. In Section 6, we provide estimated effects for this cohort from two separate achievement follow-ups, one prior to the intervention modifications and another after the full two years of the program. We further provide effects for the cohort that was subject to the intervention starting in the third term grade 2, stopped being directly subject to the interventions at the end of grade 3, and were tested again at the end of grade 4, one full year after leaving the program.

3.2 Implementation of Treatments

To implement the treatments, this program consisted of materials, personnel, and training, almost of which was implemented exclusively through the existing government system.

Ghana Education Services (GES), an agency of the Ministry of Education, designed the materials with inspiration from materials previously developed for use by Pratham for targeted and remedial instruction in India. A team of government officials from Ghana traveled to India to learn about their materials and approach. A TCAI technical assistance unit consisting of 4 Regional Coordinators provided limited support to the material development team. Teams developed specific materials for each remedial learning level, focusing on basic reading and computation skills. Remedial lessons involved outlines of topics to cover, but were not scripted lessons. Broader materials were developed to increase child focused learning across all treatments, including a bank of fun, child focused potential activities. Treatment assistants and teachers were responsible for their own lesson plans with the provided materials as a guide.

Schools in all three assistant-based treatments received the same hiring instructions. School Management Committees (SMCs) and Parent Teacher Associations (PTAs) were to identify potential assistants to be interviewed by a panel of local, GES, and NYEP repre-

sentatives.[14] Assistants were to be aged 18 to 35 and to have completed secondary school with passing grades in math, English, and science. They were also expected to speak, read, and write the school's NALAP language and live in the school community. Most, but not all, TCAs fit these criteria. See the Section 5.3 for additional discussion and tests showing balance across treatment arms.

Oversight of schools continued as usual with the addition of the 4 TCAI Regional Coordinators, each responsible for 100 schools. Existing Circuit Supervisors continued their scheduled visits of schools, which should have led them to visit each of their schools 3 to 4 times per month but their actual visits were likely less frequent. They were not specifically trained or engaged in the treatment. The specific TCAI Regional Coordinators provided at most limited support to individual schools and primarily focused on ensuring material distribution.

The TCAI intervention occurred during three academic years. Initial trainings occurred in May (Term 3) of the 2010-2011 academic year. Remedial, review, and targeted instruction lessons were to start immediately. An additional training to refresh prior participants and train additional participants occurred prior to the start of term 1 of the 2011-2012 school year. The training that modified the remedial and teacher-led interventions occurred at the start of the second term of the 2011-2012 school year.[15]

Material delivery was to occur immediately after the first training, but experienced substantial delays. Only 12 percent of head teachers surveyed at the end of the first term of implementation had received materials. By the third term of implementation (term 2 of the 2011-2012, second, academic year), over 90 percent reported having received materials. The labels above the line in Figure 1 display the academic year and intervention timeline. The labels below the line are the nine data collection points.

[14]GES designed the interview process to reduce the likelihood that the assistant was recruited on a political basis. In our sample, 94 percent of schools had a SMC, 98 percent had a PTA, and 99 percent had at least one.

[15]Teachers or assistants who joined schools between scheduled trainings received school-based orientation and training from their colleagues and then joined in the next scheduled training.

[Figure 1 about here]

4 Empirical Strategy

The primary conceptual difficulty in assessing the effects of various inputs into the education production function is the typical non-random allocation of resources and their correlation with household and school attributes, leading to biased estimates of their effects on learning. To alleviate this concern, we designed a randomized controlled trial of our four interventions.

We randomly divided our 500 school study sample schools into five treatment groups– assistant-led remedial instruction during school, assistant led remedial instruction after school, assistant split, teacher-led targeted instruction, and control.

From this randomization design, comparing outcomes between students in the treatment and control schools is straightforward. Formally, we estimate an intention to treat specification

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

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

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. Even though the assistant-led remedial interventions targeted remedial learners, we include all students in our

estimates of the effect sizes with additional analysis for heterogeneity by baseline test score and score relative to the school's score distribution.

When considering outcomes for teachers, assistants, or head teachers, we modify the i in Equation 1 to be the person of interest and adjust the outcome accordingly. For teachers, we test attendance and time on task. For assistants, we test for differences in implementation across the three assistant arms. We are unable to test assistant outcomes relative to the control and teacher-led targeted instruction arm as they did not have assistants. For head teachers, we test the effect on their attendance.

5 Sample Selection and Data

In this section we first describe the randomization procedure and then the data collected.

5.1 Sample Selection

We started with the universe of government primary schools from the official Education Management Information Systems (EMIS) school list. Selecting the exact schools to participate in the study was done in two stages. First, 42 districts were selected from 168 of the 170 districts in Ghana, ensuring at least two districts from each of the 10 regions.[16] At the district level, we randomized whether we would select 11 or 12 schools from that district. Within each district, we divided schools by whether the school was urban according to the EMIS. When possible, an equal number of urban and rural schools were selected within each district.[17]

Once the 500 school sample was selected, schools were randomly allocated into the four treatment arms and control group, stratified by region, average baseline student test score above/below median, and pupil teacher ratio above/below median. At the school level, a

[16] GES requested two districts be excluded due of issues related to the NALAP.

[17] Three districts did not have enough rural schools, and therefore have more urban schools in the sample. Twenty-one districts did not have enough urban schools so these districts have more rural schools than urban schools in the sample.

maximum of 25 pupils from each grade 1 through 3 were randomly selected using the class registers collected during a pre-baseline school survey. When grades had fewer than 25 pupils, all pupils were interviewed.[18] If a selected student was absent during the baseline visit, that student was replaced in the sample (if the class had more than 25 pupils). When possible, an equal number of male and female students were selected.[19]

This study starts with students in grades 1 through 3. We focus on students who were in grades 1 and 2 at the time of the baseline and would have been in grades 3 and 4 at the second follow-up if they continued school on pace.[20]

5.2 Data Collection

To evaluate the effect of the four models of targeted instruction we collected nine separate rounds of data between October 2010 and July 2013. In this sub-section we describe the data collection rounds. Figure 1 displays the academic year, intervention, and data collection timeline. In the next sub-section, 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 through 3 students were interviewed. We tested selected students using bespoke exams that we developed in collaboration with the Assessment Services Unit of the Curriculum Research and Development Division of GES. We further had the support of a psychometrician piloted the exams on 300 students to test for validity and reliability. Students were tested in English, math,

[18] In two schools, where class registers were not available and the class enrollment was above 25, a numbered list of all pupils in the class was created at the time of the survey with the help of the Head Teachers, and an enumerator used a table of random numbers to do the selection.

[19] Due to budget constraints, in the second follow-up the sample for the oral test was reduced to twelve for students who were grade 2 at baseline and should have been in grade 4 during this follow-up. Students from the initial list were randomly selected to take the oral portion of the test.

[20] We do not use the data from the grade 3 students at baseline because they aged out of the program after one term of implementation, a term in which only 12 percent of treatment schools had materials. As originally designed, implementation should have started earlier in the school year, exposing these students to a longer treatment before aging out.

and the school's NALAP language. The tests covered the critical objectives of the official curriculum for grades 1-3, covering a range of skills, beyond what the targeted instruction or remedial materials covered. One common test was used to test pupils across all grades.

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 but did not follow students beyond expected grade 4. In follow-up 1, our baseline students who started in grades 1 and 2 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. The data collection strategy included testing students in school and tracking those children who were not at school on that day. The survey teams attempted to track all students not in school to their homes or new residence if they had migrated.

Exams in each round where similar in spirit but contained different questions. The follow-up exams included additional written grade-level specific components. We use item response theory within each round to solve for students' latent knowledge and standardize these scores based on the control group students' mean and standard deviation.

5.3 Summary Statistics and Baseline Balance

Students

Table 2 contains summary statistics and baseline balance checks for grade 1 and 2 students surveyed and tested at baseline. 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. In all cases, we fail to reject the null hypothesis that the means are equal. The test scores are adjusted using item response theory and then standardized based on the control group scores. About 30 percent of the students wore clean, good quality uniforms and over 90 percent wore shoes. On average students were about 8.4 years old. When disaggregated by grade, the grade 1 students were on average 7.8 and grade 2 students 9.0 years old. If students started school on time at age 6 and did not repeat any grades, then they would be 6 to 7 years old in grade 1 and 7 to 8 years old in grade 2. Therefore, many of our students likely either repeated a grade or started grade 1 late. About half of the sample has a literate father and about a third has a literate mother. Therefore, many of these students are likely first generation learners. Nevertheless, about two-thirds report that they have someone at home who can help them with their homework. Students report being absent about 0.8 days in the previous week.

[Table 2 about here]

Part of the motivation for this study was that students were both behind grade level and exhibited substantial heterogeneity within classrooms. By the end of the first term of grade 1, fewer than 40 percent of grade 1 students could correctly select an upper case letter from among a list of four English letters. Among grade 3 students, 13 percent could not select the

correct English letter. Overall, on exams designed to test grade 1 through 3 materials, grade 3 students on average answered 31 percent of the English and 46 percent of both the local language and math questions correctly. Figure 2 shows both the standardized average test score by grade level and the average gap between the 90th and 10th percentile student within each school at baseline. Both student achievement and heterogeneity increase with grade level. At each grade level the average student's test score was about 0.7 standard deviations higher than the previous grade level (solid blue line). The within school achievement gap is about twice this size in grade 1, increasing to about 2.5 times this gap in grade 3 (dashed red line).

[Figure 2 about here]

Teachers

Panel A of Table 3 contains summary statistics and baseline balance checks for teachers surveyed at baseline. As with Table 2, columns 1-5 contain the means and standard deviations and column 6 the test of equality. In all cases we fail to reject equality across the treatment arms.

Just over half of the teachers are female, and they are on average about 36 years old. About 60 percent live in the community in which they teach and have on average about 10 years of experience as teachers. Around 85 percent were employed directly by Ghana Education Services, indicating that they were permanent teachers.[21]

[Table 3 about here]

Schools

Summary statistics and baseline balance checks collected at the school level appear in Panel B of Table 3. 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 is about 119, or about 37 students per grade. On average about 3.5 teachers

[21]The other 15 percent were employed by NYEP, NSS, the community, or an NGO or were volunteers.

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

Table 4 contains the demographic characteristics of assistants.[22] 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 TCAI 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 TCAI. Over half had some teaching experience. This experience including tutoring, teaching in private schools, and teaching in government schools.

[Table 4 about here]

According to the instructions given to the communities, all assistants should have been interviewed, been asked to present evidence that they passed the high school exit exam, completed high school, and been able to read, write and speak the school's official local (NALAP) language. According to the assistants these instructions were largely followed. Based on self-reports almost three-quarters were interviewed, about 65 percent were asked about their exit exam scores or whether they passed the high school exit exam, and over 90 percent were able to read, write, and speak the NALAP language, were able to speak the

[22]The assistant demographic data were collected during the spot-check data collection rounds because assistants had not been hired prior to the baseline.

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.

6 Results

In this section we provide estimates of the effect of the program on achievement, selection into the test, other student outcomes, program implementation, and time on task. We conclude this section by testing for heterogeneity by baseline characteristics.

6.1 Achievement and Selection into the Test

Achievement

Table 5 contains the results of the estimate of the effects of the four treatments on student test scores collected in the two achievement rounds, an estimation of Equation 1 with a student test score as the outcome of interest.

For each test score, we report the results from both the first and second follow-up. Recall that follow-up 1 occurred near the start of the second academic year, about a term and a half after implementation. The timing of this follow-up as an assessment of the impact of the program was not ideal for two reasons. First, students had returned from their summer holidays only about a month and a half before the surveys began. Second, while not known at the time, the model of the intervention was about to change based on feedback collected from the teachers and assistants. Follow-up 2 occurred about two years after implementation, near the end of the third academic year.

The sample for this table is students who were grade 1 at baseline. These students should have been in grade 2 in the first follow-up and grade 3 in the second follow-up, if students were progressing apace. We interviewed and assessed them regardless of their grade-level at follow-up. In the follow-up surveys students completed oral exams on grade level 1 through

3 content and written exams covering primarily grade 2 content in follow-up 1 and grade 3 content in follow-up 2. This table uses the entire test, with achievement converted to latent scores using item response theory and standardized using the control group mean and standard deviation.[23]

[Table 5 about here]

The first two columns contain the results for the combined English and Math score. Even after only about a term and a half of treatment, students' test scores were about 10 percent of a standard deviation larger for the two assistant-led remedial interventions (column 1). The other two interventions had smaller, statistically insignificant test score gains. Below the coefficient estimates we test for the equality of the coefficient of each intervention relative to the other three interventions. We fail to reject that all interventions had the same effect. In follow-up 2, test scores increased across all of the interventions by 0.08 SD (teacher-led targeted instruction) to 0.15 SD (assistant-led remedial after school) (column 2). As with follow-up 1, we fail to reject equality between the coefficients. Between follow–up 1 and follow-up 2, students were both exposed for a longer duration and the precise implementation of the targeted versions changed.

The remaining columns of Table 5 display the effects separately for English (columns 3 and 4), math (columns 5 and 6), and local language (columns 7 and 8). Odd numbered columns are from follow-up 1 and even numbered columns from follow-up 2. The English and math results are similar to the combined scores in point values and statistical significance. For local language, the results are only statistically significant for the assistant-led program during school. Among all the pair-wise tests, we only find statistical differences between interventions when comparing the assistant-led remedial after school model to either the assistant split or teacher-led model for the follow-up 2 English scores. Figure 3 graphically presents the results from Table 5.

[23]To ensure comparability between this table and the table focusing on students who were grade 2 at baseline, we standardize based on the combined mean and standard deviation across the two grades.

[Figure 3 about here]

To place these test score gains relative to a year of learning in this context we compare the test scores of students in grades 2 and 3 in control schools during follow-up 2. Grade 3 students in control schools scored 0.41 SD higher than grade 2 students.[24] Therefore, the test score gains from the treatments were equivalent to an additional 18 percent (teacher-led) to 34 percent (assistant led after school) of a grade level.

The largest point values occurred in the interventions in which assistants worked with remedial learners. In almost all cases, the point values of the during-school and after-school interventions were very similar. Therefore, students did not appear to be at a disadvantage on grade-level content when they were removed from classrooms to receive remedial attention, potentially because the classroom content was too far from their learning level, teachers were not engaged in teaching (in our sample teachers were only engaged with learners 36 percent of the time), or even when engaged with learners teachers were not using effective pedagogy. Both the assistant-split and teacher-led targeted instruction interventions tend to have smaller point values. The assistant split combined the active pedagogy with a smaller class size that focused on grade-level content. While the pedagogy could have been more engaging to the learners, and was provided in a smaller setting, it was likely still at a level too high for many to understand. Relative to the assistants, the teachers had the lowest fidelity of implementation–they were dividing students by learning levels only 5 percent of the time relative to the average assistant meeting occurring 29 percent of the time. Even though teachers were not dividing their students by learning level, they did increase their use of materials by 400 percent and their likelihood of being engaged with students by 31 percent. Therefore, students received more content, just potentially not in the full manner intended by the design of the program.

To make our results comparable to the existing studies in India, we calculate results based on the sub-set of questions most similar to the annual status of education report (ASER)

[24]This gap is smaller than the grade level gap at baseline. Not all baseline students were present at the follow-up and this exam included an additional written component.

questions used to level and assess students in India.[25] These questions focus on foundational content in both English and math. These results appear in Appendix Table A2.

When considering only the foundational content, the point values are very similar to the overall results for the two assistant remedial models. The point values are smaller in magnitude for the assistant split, likely because the assistant-split focused on grade-level and not foundational content. The point values are larger in magnitude for the teacher-led model, but still smaller than either of the assistant models. When considering only the foundational questions, the assistant-led remedial after school model is statistically different than the assistant-split model overall and when only considering English.

In contrast to the grade level results, the three models with a focus on remedial skills statistically significantly improved the foundational local language test scores. Given the delays and complications with the NALAP program, grade-level local language content could have been too difficult for all students.

Short-term vs. Persistent Achievement

The results thus far have focused on students who were grade 1 at baseline and were expected to be in grade 3 at follow-up 2. A cohort one year older was similarly tracked from the baseline through the second follow-up. These students started the intervention at the end of grade 2, exited the program at the end of grade 3, and were tested near the end of grade 4, if students progressed apace. One of the theories underpinning targeted instruction is that once students learn the foundational material, they will grasp material in subsequent grades more easily.

Table 6 contains results for this older cohort. These students are different than the students who were in grade 1 at baseline (and expected grade 3 at follow-up 2) in three important respects–they were exposed for a shorter duration (1.3 years vs. 2 years), at least half of their exposure was prior to the reformulation of the implementation, and they were

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

tested a full year after leaving the program. These students completed the same oral portion of the test as the younger cohort and a harder written portion focused on grade 4 material.

While the point values for all four interventions are positive in Table 6, the effects are only statistically different from 0 for the assistant led remedial after school (0.08 SD) and the assistant split (0.12 SD). The assistant split and teacher-led targeted are statistically different. Of particular note is that the largest point value, which is statistically different from teacher-led, is for the intervention that focused on grade 3 content. This focus on grade level materials could have prepared students more for grade 4 content than the remedial interventions.

[Table 6 about here]

Appendix Table A3 presents the results when considering only the foundational concepts for this cohort. When considering the foundational concepts, all three assistant led models show persistent, positive effect sizes, ranging from 0.11 (assistant led remedial, during school) to 0.13 (assistant split) standard deviations. The teacher-led model is no longer statistically significant.

Therefore, three of the four interventions show persistent effects on foundational content even for students who are one year out of the program. The grade-level results are not as strong, with statistically significant results for only two of the four interventions, providing some evidence that a strong foundation persists into material that includes grade level content.

Selection Into the Test

Even though we demonstrated baseline balance in observables, an additional concern is that differential student attrition led to an imbalanced sample during the achievement tests. To minimize student attrition, students were tracked, if possible, and asked to come to school to take the exam. Nevertheless, not all students from the baseline took the two achievement exams. In Table 7 columns 1 and 3, we test for differential selection by treatment status at the two follow-up rounds. About 78 percent of control group students who were in grade 1

at baseline took the first follow-up exam and 73 percent took the follow-up 2 exam. We find no statistically significant selection into either testing round by treatment status (columns 1 and 3) but fail to reject that the treatment effects are jointly 0 for the first follow-up (p-value = 0.07, column 1). We further test for differential selection by treatment status and baseline test score (columns 2 and 4), and find no evidence of statistically significant differential selection into either round of tests. Further note the low R-squared–treatment status and the other covariates explain very little of the variation in the likelihood that a student is present in either round. Even though the difference between the different treatments status are at most minimal, we calculate Lee (2009) bounds and find similar achievement effects with the adjusted levels of attrition. Those results appear in Appendix Table A4.

[Table 7 about here]

6.2 Other Student Outcomes

The child-centered learning promoted by the TCAI teaching and learning materials could have made school more enjoyable causing students to be more likely to attend and less likely to drop out. In the first two columns of Table 8 we test for these outcomes using data collected during the unannounced school visits.

[Table 8 about here]

We measure student attendance from 0 to 1, averaged over all visits to the students' schools. Overall, student attendance is low with the average control group student present only 64 percent of the time (column 1). The interventions did not change this likelihood with small (less than 1 percentage point) and statistically insignificant point values. Students are similarly no more likely to have stopped attending this particular school entirely (column 2).[26]

[26]Both students who changed schools and those no longer attending any school are registered as 0 in column 2. The sample size changes across columns as not all questions were asked in all rounds and teachers did not respond about all students. The results are similar when the estimates are limited to the same sample.

One unintended consequence of our program could have been that teachers were more aware of student learning levels due to their termly assessments and therefore might have been more likely to encourage students who were behind to repeat grades. The intervention did not have this unintended consequence as the likelihood of demotion from the expected grade was 18 percent in the control group and not statistically different for the treatment groups (column 3).

6.3 Program Implementation and Time on Task

Teachers and Head Teachers

Just as the program could have made the school day more enjoyable for students, it could have made teachers more likely to attend school. In contrast, the presence of an assistant to cover classes for absent teachers could have caused teachers in the assistant-focused interventions to be less likely to attend. At each spot check enumerators asked head teachers who the primary classroom teachers were for grades 1 through 3. Enumerators then recorded whether the teacher was present on the school grounds. In column 1 of Table 9, we find that on average only 69 percent of teachers in the control group were present at the start of our unannounced checks with no statistically significant differences by treatment status.

[Table 9 about here]

Even though teachers were no more likely to be at school, teachers in the teacher-led targeted instruction intervention increased their time on task. Teachers in the teacher-led targeted instruction intervention were 5 percentage points more likely to be in their classroom (column 2), relative to a control group mean of 52 percent. They were also 11 percentage points more likely to be engaged with students, a 30 percent increase over the control group mean of 36 percent (column 3). We reject that this effect is the same across all treatment arms (p-value=0.00).

As part of the intervention treatment teachers and assistants received teaching materials. These could have been a complement to or a substitute for other teacher-made teaching

materials. Each assistant was assigned a mentor teacher and this teacher or other teachers in the school could have borrowed the TCAI materials or been inspired to make their own. We recorded material use for classrooms in which a teacher was present. Only about 5 percent of control group teachers were using teaching and learning materials and teachers in the assistant-focused interventions were no different (column 4). Teachers in the teacher-led model were 21 percentage points more likely to be using materials, a 400 percent increase over the control group.[27] The assistant-focused interventions did not cause the regular classroom teachers to make or use their own materials.

For the teacher-led arm only, in the final two spot-checks, enumerators recorded whether teachers had divided their students across grade levels by learning levels. Of those teachers who were engaged with students, only 5% had split their classrooms by learning levels.

We also recorded whether head teachers were present upon our arrival. Head teachers in the assistant-led remedial during school and the teacher-led targeted instruction model were more likely to be present at school by 14 and 9 percentage points, respectively, relative to control group mean of 51 percent (column 5).

This increase in teacher time on task and head teacher attendance occurred with no change in incentives and minimal increases in monitoring relative to the control group and no monitoring difference between the groups. For teacher attendance and presence in the classroom and head teacher attendance, we fail to reject equivalence across the treatment arms. For teachers being engaged with students, we reject equality of the coefficients across the four arms.

Assistants

In Table 10 we test for differences between the three arms that included the assistants and the evolution of assistant behavior over time. In each column, the outcome is an indicator for whether the activity occurred. As only three treatment arms had assistants, the treatment

[27]We separately test for whether the intervention increased the likelihood of teachers having teacher made materials visible (11 percentage point increase in the teacher-led arm) or in use (no statistical difference) or having TCAI materials visible (32 percentage point increase in the teacher-led arm) or in use (20 percentage point increase in the teacher-led arm).

coefficients are relative to the assistant-split. The mean for the assistant-split treatment appears at the bottom of each column. We also provide the activities observed in years 2 and 3 of the intervention relative to year 1. The year 1 average also appears at the bottom of each column.

[Table 10 about here]

Average assistant attendance was lower than for teachers–only 63 percent of assistant-split assistants were present (column 1). This value is statistically the same for the during school assistant-led remedial model and 16 percentage points lower for the after-school remedial split, potentially because the assistants were not required to be present during the school day when our enumeration teams arrived. Over time assistant attendance decreased with maximum attendance happening in Year 1 (69 percent), a 13 percentage point decrease in Year 2, and a 27 percentage point decrease in Year 3. By Year 3 the average attendance was only 43 percent, about two-thirds the attendance rate of teachers.

The assistants in the assistant-split intervention were the most likely to be covering for a classroom teacher–14 percent on average, 13 percentage points higher than the after school model and 8 percentage points more likely than the assistant-led remedial during school (column 2). While teachers might have learned over time that they could use these assistants to replace themselves, the year 2 and 3 coefficients are statistically equal to each other and to year 1. The desire to increasingly rely on the assistants over time could have been offset by the increasing frequency of their absence.

The assistants were holding their smaller group meetings 23 percent of the time in the assistant-split model, 28 percent of the time (not statistically different) in the remedial during school model, and 36 percent of the time in the after school model (statistically significantly different from the other two) (column 3). The likelihood of the meetings decreased over time with it happening 40 percent of the time in Year 1 with 13 and 21 percentage point decreases in years 2 and 3, similar magnitudes to the increased absenteeism in those years.

Almost all of the assistants divided their students in the manner in which they were

supposed to–either randomly or by learning level. The during-school remedial model was 7 percentage points more likely than the assistant-split model to use the correct method. The likelihood of this happening decreased by 13 percentage points in year 2 and 7 percentage points in year 3. For those assistants who did not use the correct model, divisions were most likely to be done based on teacher or head teacher suggestion or arbitrarily. In some cases students were never split at all.

Across all three arms, assistants were equally likely to have the TCAI materials visible and be using them (about 40 percent, columns 5 and 6). The use increased over time. Around 20 percent of assistants had TCAI materials and were using them in Year 1, increasing by about 18 percentage points for both in year 2, and 50 percentage points for visibility and 37 percentage points for use in Year 3.

6.4 Heterogeneity

The analysis thus far focused on the test scores of all students as even the remedial interventions could have helped all students. The during school assistant-led remedial model created a smaller, more homogeneous classroom for those students who were not working with the assistant. The after-school assistant-led remedial model, by increasing the learning levels of the remedial students, could have created a more homogeneous learning environment during the school day with less disruption as more students could engage with grade-level content. Nevertheless, part of the motivation of the study was the potential for all of these interventions to increase student test scores at the lower end of the test score distribution.

In columns 1 and 2 of Table 11 we test for heterogeneity by two measures of student baseline achievement–student test scores and whether the student was in the bottom third of his or her grade by school test score distribution.

[Table 11 about here]

In column 1 we interact baseline test score with each of the treatment indicators and find no statistically significant evidence of heterogeneity. While imprecisely measured, the

positive coefficients on the assistant-led remedial during school and the assistant-split interventions show relatively higher scoring students benefiting more. The negative coefficient on both the assistant-led remedial after school and the teacher-led interventions, also imprecisely measured, show relatively weaker students benefiting more.

In column 2 we interact whether the student scored in the bottom third of the school by grade baseline score distribution with each treatment, a coarse approximation of remedial status.[28] We continue to control for baseline test scores and include an indicator for being in the bottom third of the school by grade distribution. In all cases the coefficients are negative–students in the lower third benefited less conditional on their baseline scores. The interaction is only statistically significant for the assistant split and is of larger magnitude than the main effect–in expectation students in the lower third of the score distribution did not benefit at all from the assistant split that only reviewed grade level material, likely because these students' learning levels were well below the target ability for the grade level material.

In column 3 of Table 11 we test for heterogeneity based on student gender by interacting each of the four treatment variables with female. In all cases, the interaction term is positive and in three cases statistically significant–female test scores increased more than male test scores in the three interventions that had some remedial focus. In all cases we reject that the coefficient on the sum of the interaction and the main effect sum to 0 indicating female students scores statistically significantly increased across all interventions. The main effects, i.e. the effect on male students, are now statistically insignificant for both the assistant-split and the teacher-led targeted instruction model.[29] These differences between genders are likely not due to role model effects as a larger percentage of teachers than assistants were

[28]In the rounds in which remedial was measured, it was only measured for the two assistant-led remedial interventions. Approximately one third of each grade level flagged as remedial was considered remedial in the data collected.

[29]We find no evidence of heterogeneity by pupil teacher ratio in lower primary grades, whether schools used multigrade classrooms in lower primary grades, whether the school was classified as rural, or whether the school was in a location classified as deprived. In our sample, pupil-teacher ratio ranged from 4.5 to 1 to 150 to 1 and 18 percent of students were in schools with multi-grade teaching, 65 percent in schools classified as rural, and 38 percent in schools in locations classified as deprived.

female (53 percent of teachers and 43 percent of assistants).

7 Implementation Challenges and External Validity

We first discuss the various challenges encountered in implementing a program almost entirely within existing government systems and then discuss how our implementation findings are similar, and different, from those in other contexts.

7.1 Implementation Challenges

All four interventions faced implementation difficulties related to the challenges in scaling a program within existing government structures.

Across all arms, materials were delayed in reaching schools. Most schools did not receive their materials until one to two full terms after the training.

Both teachers and assistants were often absent. Across all three assistant arms, we found assistants present 56 percent of the time, with a low of 43 percent in year 3. Some of their absenteeism was likely related to their lack of payment. NYEP was responsible for their payments, and these were often delayed, including a span of 8 months when assistants received no payment. During the year 2 follow-ups, some schools reported that their assistants were striking due to non-payment of salaries. Across all arms, teachers were absent about 30 percent of the time, causing assistants to fill-in for teachers or making the teacher-led targeted instruction model less useful than it could have been.

Teachers and assistants were only subject to existing, system-wide support and incentives. We did provide one Regional Coordinator per 100 schools to oversee all operations in those schools. These individuals were mostly involved with material delivery and solving logistics problems, not supervising or supporting teachers or assistants. This is in contrast to creating a separate supervisory layer for teachers in Haryana, India or the oversight and support provided to the assistants in the models in Banerjee et al. (2007), Banerjee et al. (2010),

and Banerjee et al. (2017). Because of the nature of the assistants and their expected future relationships with the schools, they faced fewer incentives than the contract teachers in Duflo, Dupas, and Kremer (2015). Fewer than half of our assistants aspired to teach in the future, and for those who did, they would have had to enroll in additional courses to earn certification.

An additional challenge was the model change for the assistant-remedial and teacher-led targeted instruction part way through the implementation. After about two terms of implementation, at the start of the second term of the second year, the assistants switched to focusing on one learning level at a time–working first to bring Level 1 students to Level 2–and teachers no longer worked within their grade-levels, instead dividing students by learning level across grades.

7.2 External Validity

Despite these challenges, recall from Table 5 and Figure 3 that each intervention had statistically significant, positive effects on student achievement by the second follow-up. Figure 4 places the effect sizes in context with the existing evidence on similar interventions in India and the class size reduction and student tracking interventions in Kenya. The first four solid bars reproduce the two year effects from Table 4 column 2.[30] The remaining nine striped bars are the combined language and math test score effects from other studies.[31]

Upward sloping diagonals are from India, downward sloping diagonals are from Western Province, Kenya. Red bars involved additional staff added to the school, whether assistants (this study and India) or contract teachers (Kenya). Blue bars focused on teachers. Purple bars included both teacher training and additional staff.

Relative to the other interventions that added staff to the system (the red bars) our point

[30]Since our model evolved and improved over time, we present the two year effects. Where possible we present the two year effects from other studies as well.

[31]If a study only reported separate language and math effects, then the effect presented is the average of the two effects. The separate point estimates and additional details about each study appear in Appendix Table 1.

estimates are smaller, perhaps because this implementation lacked the NGO supervision that occurred in India and assistants were paid less consistently than in Kenya.

In contrast to two of the three previous interventions on teacher-led targeted instruction, our teacher-led targeted instruction model increased test scores despite low adherence to dividing students by learning level. Our effect sizes are a similar magnitude to the effect of the intervention that trained teachers, added an additional supervisory layer, and an NGO directly supervised teachers in Haryana (Banerjee et al. 2016).[32]

Given the challenges relative to the more ideal models implemented in India, the positive effects attest to the commonality between contexts of the binding constraints on effective teaching. Adherence to a model closer to that achieved in India could result in even larger effect sizes.

8 Cost Effectiveness

We present a conservative estimate of the costs of the program–providing the costs of the program using the ingredients method as the program was designed, not as it was imperfectly implemented. We fully cost all payments to assistants each year, even though they were not always paid on time. We also cost the full year of the materials, even though materials were delayed. Based on estimates of scaling a single program to the entire country, the per student annual costs would be $19.60 for the remedial assistant interventions, $18.77 for the assistant-split, and $10.65 for teacher-led targeted instruction. When considering cost-effectiveness, we follow Kremer et al. (2013) and put each intervention on effect size per $100 scale. Our students received effectively two years of the intervention, spread across three academic years. The effect sizes per $100 are 0.21SD for the assistant split, 0.36SD for the teacher-led targeted instruction, and 0.38 SD for the assistant-led remedial during school and assistant-led remedial after school. The similarity of the point estimates of the

[32]The study in Haryana found positive statistically significant effects on language tests and null effects on math. This bar is the average.

three targeted instruction or remedial interventions are remarkable–the assistant-led ones cost approximately twice as much per student with approximately twice the benefit.

Because we have a multiple year intervention we can also consider the cost effectiveness of a shorter duration, i.e. lower dosage, of the program. The first follow-up occurred near the start of the second school year, a little over one term into the program. As the effect sizes for the combined English and Math score at this first follow-up is between 53 and 69 percent of the point values for the second follow-up and the costs were less than half, a shorter dose of the program appears to be more cost effective.

9 Conclusions

Ghana, as many other low-income countries, has largely eliminated the school-based barriers to primary school enrollment, but now faces the dual challenge of low average student achievement and heterogeneous classrooms. Building on evidence from seven separate interventions in India and two studies in Kenya, we worked within existing government structures to use an RCT to simultaneously test four interventions to improve student achievement in lower primary schools 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 teacher's aides. The assistants either operated a remedial pull-out program (assistant-led remedial instruction during school), provided after school lessons for remedial learners (assistant-led remedial after-school), or divided the learners between the teacher and themselves for part of the school day (assistant-split). The final intervention used existing teachers who were instructed to divided three grade-levels of students by learning level instead of grade-level for a part of each day.

All four interventions increased student learning based on a combined written and oral test administered at the end of (expected) grade 3 for those students who started the program in grade 1. Effect sizes range from 0.08 (teacher-led targeted instruction) to 0.15 (assistant-led remedial after school) standard deviations. Students who were exposed to the program

starting in grade 2 and tested at the end of grade 4, one year after ending the program, the effect sizes are smaller in magnitude (with the exception of the assistant split) and the teacher-led model is no longer statistically significant.

We find no evidence that the program increased student attendance, drop-out, or likelihood of being demoted. The teacher-led model increased the likelihood that teachers were engaged with students and using teaching and learning materials.

All models faced issues of material delays, teacher and assistant absenteeism, and weak mechanisms for support and monitoring.

When considering cost effectiveness, the assistant-led after school remedial program, assistant-led during school remedial program, and teacher-led targeted instruction program are similarly cost effective–the effect sizes and costs of the first two are approximately twice the size of the third.

Future research will focus on how to most efficiently and effectively strengthen government implementation systems, providing an enabling environment to implement programs focused on quality learning.

References

Andrabi, T., J. Das, A. I. Khwaja, T. Vishwanath, and T. Zajonc (2007). Learning and educational achievements in punjab schools (leaps): Insights to inform the education policy debate. *World Bank, Washington, DC*.

Bando, R., E. Naslund-Hadley, and P. Gertler (2019, September). Effect of inquiry and problem based pedagogy on learning: Evidence from 10 field experiments in four countries. Working Paper 26280, National Bureau of Economic Research.

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., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton (2016, October). Mainstreaming an effective intervention: Evidence from randomized evaluations of teaching at the right level in india. Working Paper 22746, National Bureau of Economic Research.

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

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

Beg, S. and A. Lucas (2019). Screen time: Tablets with pre-loaded textbooks did not increase learning. Working paper.

Bo, H. and S. Galiani (2019, November). Assessing external validity. Working Paper 26422, National Bureau of Economic Research.

Bold, T., D. P. Filmer, E. Molina, and J. Svensson (2019). The lost human capital: Teacher knowledge and student achievement in africa. The World Bank Policy Research Working Paper 8849.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2013). Scaling up what works: Experimental evidence on external validity in kenyan education. *Center for Global Development Working Paper* (321).

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature 48*(2), 424–55.

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., 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 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 (2018, August). Educator incentives and educational triage in rural primary schools. Working Paper 24911, National Bureau of Economic Research.

Glewwe, P., M. Kremer, and S. Moulin (2009). Many children left behind? textbooks and test scores in kenya. *American Economic Journal: Applied Economics 1*(1), 112–35.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004). Retrospective vs. prospective analyses of school inputs: the case of flipcharts in kenya. *Journal of Development Economics 74*, 251–268.

Hanushek, E. A. (2019, January). Addressing cross-national generalizability in educational impact evaluation. Working Paper 25460, National Bureau of Economic Research.

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

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher-related policies. *Annu. Rev. Econ. 6*(1), 801–825.

Kerwin, J. and R. L. Thornton (2018). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures.

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

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics 114*(2), 497–532.

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., P. J. McEwan, M. Ngware, and M. Oketch (2014). Improving early-grade literacy in east africa: Experimental evidence from kenya and uganda. *Journal of Policy Analysis and Management 33*(4), 950–976.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019, 04). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania*. *The Quarterly Journal of Economics 134*(3), 1627–1673.

Mbiti, I., M. Romero, and Y. Schipper (2019, May). Designing effective teacher performance pay programs: Experimental evidence from tanzania. Working Paper 25903, National Bureau of Economic Research.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries a meta-analysis of randomized experiments. *Review of Educational Research 85*(3), 353–394.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Muralidharan, K. (2013). Priorities for primary education policy in indias 12th five-year plan. In *India Policy Forum*, Volume 9, pp. 1–61. National Council of Applied Economic Research.

Muralidharan, K., J. Das, A. Holla, and A. Mohpal (2017). The fiscal cost of weak governance: Evidence from teacher absence in india. *Journal of Public Economics 145*, 116–135.

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

Pritchett, L. and J. Sandefur (2015). Learning from experiments when context matters. *American Economic Review 105*(5), 471–75.

Vivalt, E. (2016). How much can we generalize from impact evaluations?

10 Appendix

In the following tables we provide a number of additional estimations.

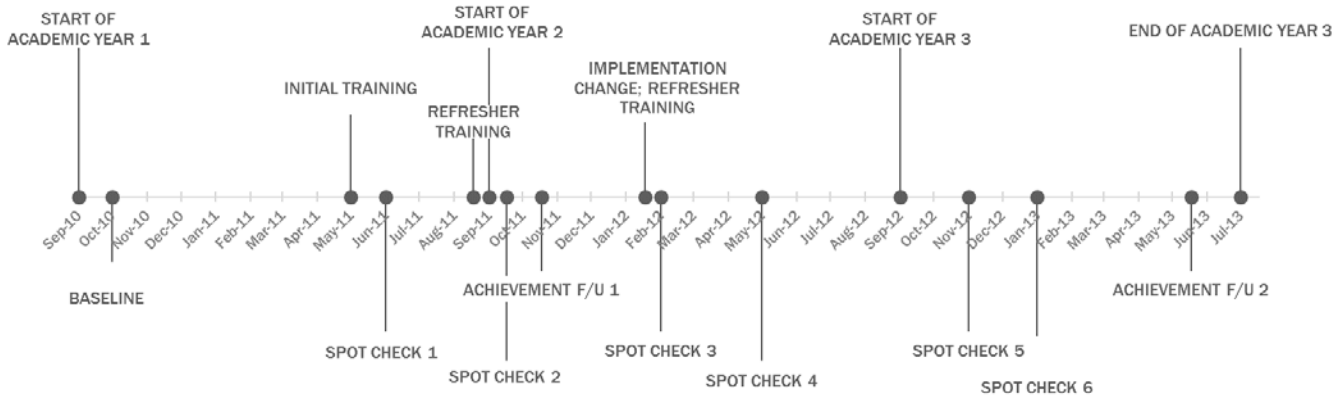
[Appendix Table A1 about here]

[Appendix Table A2 about here]

[Appendix Table A3 about here]

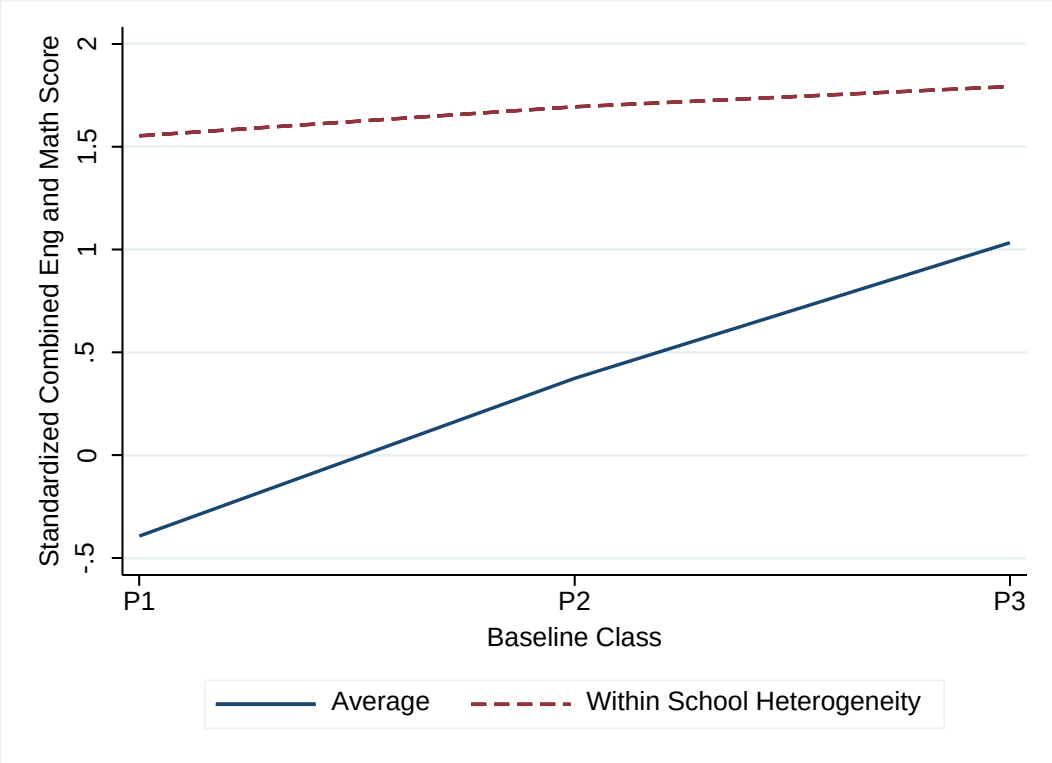
[Appendix Table A4 about here]

Figure 1: Academic Year, Implementation, and Data Collection Timeline



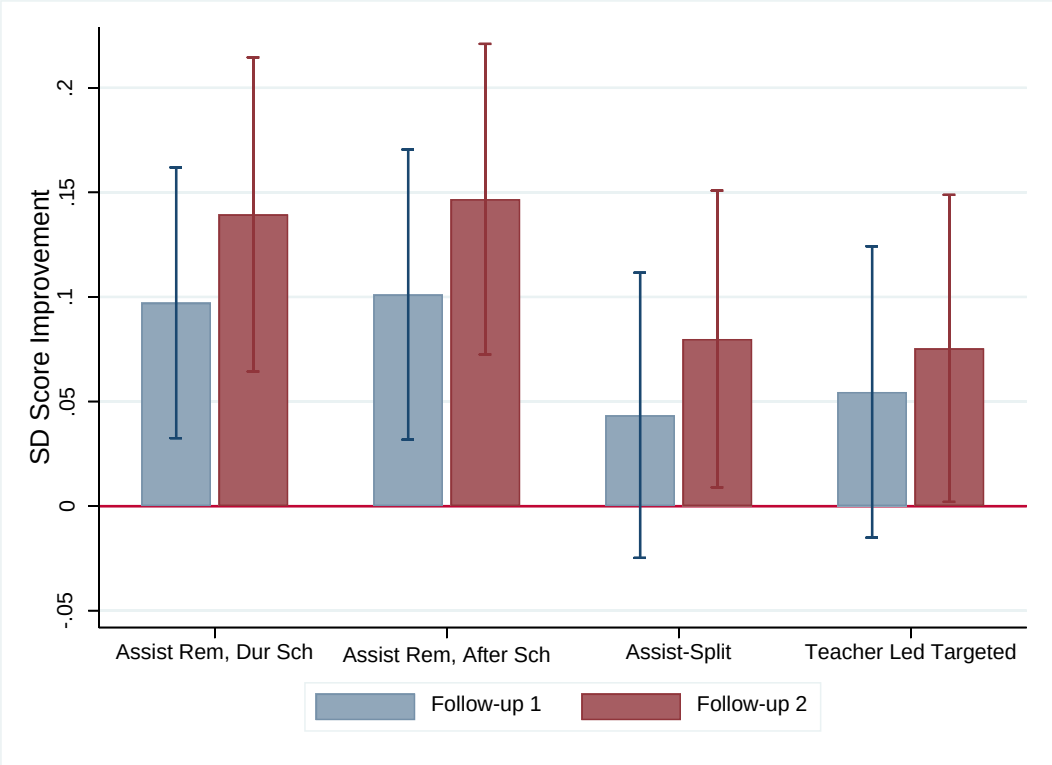
Notes: Labels above the line are academic year and implementation milestones. Those below the line are the nine data collection points.

Figure 2: Test Scores and Within School Heterogeneity



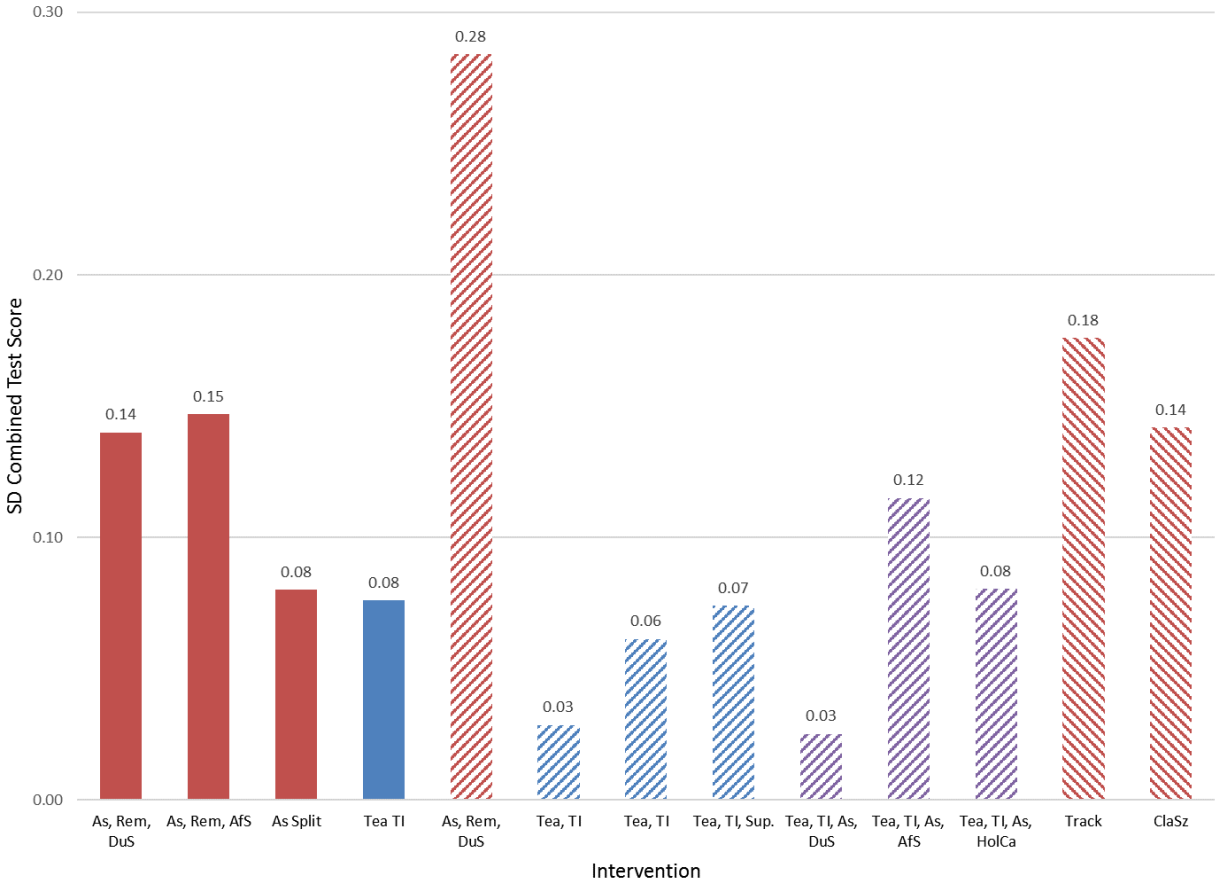
Notes: Test scores standardized to mean 0 and standard deviation 1 across grades 1 and 2. Within school heterogeneity defined as the difference between the 90th and 10th percentile scoring student within each grade by school, averaged across schools.

Figure 3: Effect Sizes from Follow-up 1 and Follow-up 2



Notes: Follow-up 1 occurred in the first term of the second academic year, the second term after the start of the intervention. Follow-up 2 occurred in the last term of the third academic year, about 2 years after the start of the intervention.

Figure 4: Comparisons across Interventions and Contexts



Notes: Solid bars are this study. Upward sloping diagonals are in India, downward sloping diagonals are in Western Province Kenya. Red bars are assistant only, blue bars are teachers only, purple bars are both assistants and teachers. The first four bars reproduce the two year effect sizes from Figure 2. As=Assistants. Rem=Remedial. DuS=during school. AfS=After school. TI=Targeted Instruction. HolCa=holiday camp. Sup=extra supervisory layer. See Appendix Table 1 and text for additional details on each intervention and the interventions included.