THE ECONOMIC JOURNAL

The Economic Journal, **134** (*July*), 1985–2008 https://doi.org/10.1093/ej/ueae003
Advance Access Publication Date: 2 February 2024

EXPERIMENTAL EVIDENCE ON FOUR POLICIES TO INCREASE LEARNING AT SCALE*

Annie Duflo, Jessica Kiessel and Adrienne M. Lucas

We partnered with the Ghanaian government to simultaneously test four methods of increasing achievement—assistant-led remedial pull-out lessons, remedial after-school lessons, smaller class sizes and teacher-implemented partial day tracking—in schools with low and heterogeneous student achievement. The interventions increased student learning by about 0.1 standard deviations, rising to 0.4 standard deviations when adjusting for imperfect implementation, with no effects on attendance, grade repetition or drop-out. Test score increases were larger for girls. Test score gains persisted after the program ended. Assistants implemented the program with higher fidelity than teachers, although their fidelity decreased over time while teacher fidelity marginally improved.

Many developing countries have eliminated the fee-based barriers to primary school enrolment, leading to a substantial rise in the number of enrolled children (Lucas and Mbiti, 2012). Unfortunately, the existing education systems, designed to serve a smaller and more homogeneous student population, struggle to accommodate more heterogeneous classrooms and first-generation learners. Effective solutions have been proposed through smaller randomised controlled trials, yet whether they can increase learning when integrated into existing systems at scale is unknown. This paper tests, in existing systems, four alternatives to support teachers' transition to the new status quo, a frontier challenge for developing countries. In a single 500-school, nationwide, randomised controlled trial (RCT) that reached over 80,000 students, we test four models that built on some of the most effective content delivery interventions in the last 20 years in developing countries—assistant teachers, smaller class sizes, additional instructional time, tracking, and remedial and differentiated instruction—and show their potential, relative effectiveness and

* Corresponding author: Adrienne M. Lucas, Department of Economics, University of Delaware, 406 Purnell Hall, Newark, DE 19716, USA. Email: alucas@udel.edu

This paper was received on 9 May 2022 and accepted on 12 January 2024. The Editor was Sule Alan.

The data and codes for this paper are available on the Journal repository (Duflo *et al.*, 2023). They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.10492707.

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

effectiveness over time when fully designed and implemented by existing government systems. Results from this study have influenced the implementation of programs to improve education in India and Africa.

The Teacher Community Assistant Initiative (TCAI) was a Ghana Ministry of Education program that implemented four interventions to increase student learning using existing schooling and youth employment systems under the unifying theory that focusing more on individual learners could improve student outcomes. In each intervention, existing education sector employees designed teaching and learning materials, trained educators in student-centred, active pedagogy and provided the educators accompanying teaching and learning materials. Three of the interventions used an existing youth employment scheme to hire teaching assistants to work with (*1*) remedial learners on a pull-out basis during the school day, i.e., pull-out remedial, (*2*) remedial learners outside of the school day, i.e., after-school remedial, or (*3*) half of the classroom each day on grade-level content, i.e., classroom split. The fourth intervention trained teachers to divide students into three learning levels for part of the day and focus instruction on students' learning levels, i.e., partial day tracking. We evaluated the effectiveness of each intervention by randomising 500 schools into one of the four treatment arms or a control group and conducting nine rounds of data collection over three school years.

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

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

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

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

In addition to already influencing policy in both Africa and South Asia, our findings make three related contributions to the economics literature. First, students not learning while in school is a primary concern in many countries, yet limited evidence exists on how to improve learning at scale within existing government systems. Our four alternatives incorporate some of the most promising findings from separate interventions into a single study within an existing system (Banerjee *et al.*, 2017; Evans and Acosta, 2021).[1] The four interventions implement instruction more focused on individual learners—during or after-school remedial lessons, by dividing the class in half, or having existing teachers specifically focus on a more homogeneous group of learners—building on Krueger (1999), Banerjee *et al.* (2007; 2010; 2017) and Duflo *et al.* (2011).[2] By comparing the effects and cost effectiveness of the four alternatives together and in a new context, we further contribute to the understanding of the external validity of these methods and which is the most effective and cost-effective way to increase learning. All four interventions increased learning, yet implementation difficulties show that the capacity of the agency in charge of implementation might matter as much as the program design.[3]

Second, this paper contributes to a broader literature on the importance of at-scale experiments implemented within existing systems. This study did not include a highly motivated Non-governmental Organization (NGO), a researcher team heavily involved in implementation, a narrowly geographically selected sample or additional personnel who were hired outside of normal government operations, unlike many experiments in development economics research (Muralidharan and Niehaus, 2017). This study instead relied on existing systems across a randomly selected nationwide sample. We show the potential for success of similar interventions at scale and highlight the additional challenges of at-scale programs.

Third, we show that existing government structures have the capacity to increase learning in spite of rigid hierarchies and wages unrelated to productivity (Muralidharan *et al.*, 2016; Bau and

[1] One potentially promising class of interventions we do not address are those using technology (see Beg *et al.*, 2022 for a summary of the literature). Requirements of security, electricity and internet connectivity rendered such interventions impractical in this context. Most education RCTs in lower-income countries only contain one treatment arm (Evans and Yuan, 2022).

[2] This class of instructional models has different names across contexts including differentiated instruction or learning, targeted instruction or teaching at the right level (TaRL). Smaller class sizes led to higher student achievement in Tennessee primary schools (Krueger, 1999). The remedial pull-out intervention was inspired by the NGO-supported assistants in Banerjee *et al.* (2007) that increased learning in Mumbai and Vadodara cities, India. The remedial after-school intervention comes from Banerjee *et al.* (2010), which increased letter recognition in Jaunpur district, India. The evidence from NGO-supported tracking programs is mixed: full-day tracking increased student learning in Western Province, Kenya (Duflo *et al.*, 2011); partial day tracking did not increase learning in Bihar and Uttarakhand states, India (Banerjee *et al.*, 2017) and partial day tracking increased learning when an extra supervisory layer and instructional hour accompanied it in Haryana state, India (Banerjee *et al.*, 2017).

[3] Based on a frequentist meta-analysis and Bayesian hierarchical models, Angrist and Meager (2022) found that much of the heterogeneity across the effect sizes of previous studies that grouped students by learning level in India can be attributed to the person delivering the program (teachers or assistants) and the fidelity of implementation.

Das, 2020). Previous programs that embedded NGO-designed programs in existing, and hesitant, government structures did not necessarily increase student learning (Banerjee *et al.*, 2017; Bold *et al.*, 2018). In this version, government involvement started at the outset in the design of the teaching, learning and training materials and continued through training and implementation, creating a truly government-owned and -operated program. The increase in test scores demonstrates the potential potency of the interventions if implemented elsewhere entirely within a government system. Yet, we also show that continuing support beyond program inception is also crucial—the assistants' adherence fell over time.

1. Background

1.1. *The Ghanaian Educational System*

Primary school in Ghana is grades 1 through 6, starts at age 6 and is tuition-free in government schools. Our study focuses on students in government schools in grades 1–3, i.e., lower primary. The school year starts in September, consists of three approximately 13 week terms and ends in July. In lower primary school, teachers are grade-level classroom teachers, teaching all subjects to their classroom of students. Teachers' salaries are paid centrally, and the Ghana Education Service (GES) assigns teachers to schools.

As with many other lower-income countries with high-stake certification exams between schooling levels, teachers are expected to adhere to a national curriculum even if students are well behind grade level. This pressure often causes them to focus on the highest-achieving students, those at grade level or above (Gilligan *et al.*, 2022). The level of the official curriculum to which teachers must adhere and pedagogical methods that teachers use are largely unchanged from a time in which only wealthier, more highly educated parents could afford to send their children to school, even though the number of children in schools and the heterogeneity of their family backgrounds and pre-school preparations have increased substantially since the start of free primary education in Ghana in 2005. This results in heterogeneous classrooms with many students left behind—only about a quarter of primary school students reach proficiency levels in English and math (Ministry of Education, 2014). In our baseline data, 94% of grade 3 students could not read a grade 3 text, 18% of grade 3 students could not identify letters of the English alphabet and the within grade-by-school heterogeneity was larger than the difference in the average test scores between grades 1 and 3.

In the year prior to the study, the language of instruction in lower primary grades changed from each school's discretion, usually a combination of English and a local language, to the school's assigned National Literacy Acceleration Program (NALAP) language. Full implementation of the NALAP program lingered into our study years (Hartwell, 2010). Because of the NALAP delays, our analysis focuses on math and English skills, providing separate estimates for NALAP test scores.[4]

[4] A school's NALAP language was determined by geography and was not necessarily the mother tongue of all or a majority of the schools' students. At baseline, 66% of control schools reported having received all NALAP materials. After-school remedial schools were statistically less likely ($p < .10$) than the control schools to have received the materials. At baseline, 72% of schools, balanced across treatment statuses, reported that their teachers were attempting to implement the NALAP program.

1.2. *National Youth Employment Program*

The National Youth Employment Program (NYEP) paid the intervention's assistants, known as teacher community assistants (TCAs). NYEP started in 2006 under the Ministry of Youth and Sports, the ministry responsible for youth empowerment, to offer unemployed youth (18–35 years old), mostly secondary school graduates, two-year public service positions and a small ($80–$100) monthly stipend, which was not always paid on time. NYEP youth were already used by the Ghana Education Service on a limited basis as community education teaching assistants to fill vacant teacher positions, often in remote areas.[5]

2. Intervention and Conceptual Framework

2.1. *Intervention*

The project was a partnership between GES, the Ghana National Association of Teachers, and NYEP. In preparation for the implementation, Ghanaian education officials visited India to learn from Pratham, a large Indian NGO, about the previous successes and challenges of the teaching-at-the-right-level (TaRL) approach that was studied by Banerjee *et al.* (2007; 2010; 2017). Government employees under the Ministry of Education umbrella designed the teaching, learning and training materials with inspiration from the TaRL approach.

This study tested four methods of improving student learning in government schools—pull-out remedial, after-school remedial, classroom split and partial day tracking—relative to each other and a control group. Treatment was assigned at the school level with 100 schools receiving each treatment. In the remainder of this subsection we first describe the common elements across the four interventions, then describe the specific details of each intervention.

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

Existing government trainers provided all educators the same training on how to engage in active, child-focused pedagogy and materials that contained suggested engaging, child-focused activities.[6] All educators were responsible for their own lesson plans with the provided materials as suggestions and guides. Students in the treatment schools were to receive the program for one hour each day, four days per week. Educators received training on how to divide the students appropriately depending on the intervention. The programs were implemented with minimal support from four regional coordinators who were each responsible for 100 regionally proximate schools and reported to the Director of Basic Education.

[5] In 2013, NYEP suffered a scandal related to financial fraud and was renamed the Ghana Youth Employment and Entrepreneurship Development Agency. In 2015, it was renamed and restructured as the Youth Employment Agency, focusing on job fairs and job search skills instead of direct employment. Assistants are not being used as part of the ongoing scale-up of the partial day tracking intervention; see Section 7.1 below for more details.

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

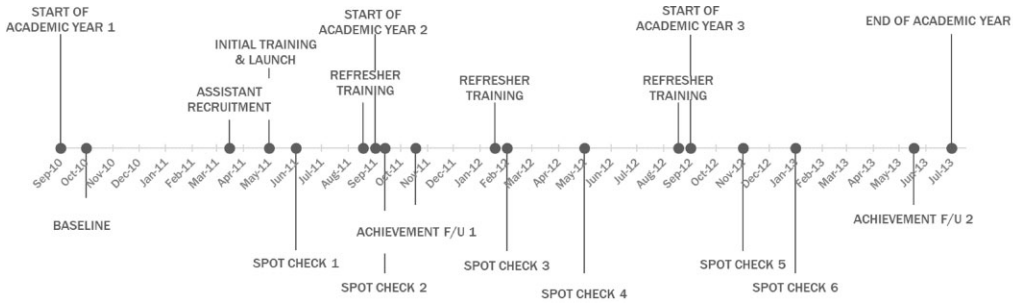

Fig. 1. *Academic Year, Implementation and Data Collection Timeline.*
Notes: Labels above the line are academic year and implementation milestones. Labels below the line are the nine data collection points.

All interventions had the same timing and implementation schedule and occurred over three academic years. Initial trainings occurred in May (term 3) of the 2010–1 academic year (academic year 1) with treatment lessons starting immediately despite material delays that lasted into the second academic year. Additional training sessions occurred throughout the next two academic years, with the study ending at the end of the 2012–3 academic year (academic year 3).

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.

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

Figure 2 summarises the components of each intervention. The interventions were not strictly nested, but did contain common elements across multiple interventions. The remainder of this subsection describes each intervention.[7]

2.1.1. *Pull-out remedial*

In the pull-out remedial intervention, assistants received additional training and materials to use with remedial learners. They tested students at the start of each term to determine whether a student was in learning level 1, 2 or 3. Learners in levels 1 and 2 were eligible for remedial lessons. Students could move between levels based on these termly exams or at any time at the assistant's discretion. The learning-level groups likely included multiple grade levels. The assistants worked with the level 1 and 2 learners separately, on a pull-out basis, one hour per day, four days per week. This gave both the remedial learners and non-remedial learners who remained in their regular classrooms more homogeneous instruction during the pull-out time. Assistants were encouraged to use extra space as available or meet with their learners outside under a tree if the school did not have extra instructional space.

[7] The selection of the interventions was a collaboration between government stakeholders and the researchers. Government stakeholders were interested in effective and cost-effective ways to increase student learning with at least one of the interventions directly involving classroom teachers to increase support from the teachers' union. The interventions further needed to work within existing systems without adding substantial staff or expenses. Education technology (Edtech) interventions were not considered due to issues of securing, powering, servicing and supporting the hardware. The three assistant interventions are similar to the group tutoring interventions highlighted by Nickow *et al.* (2020).

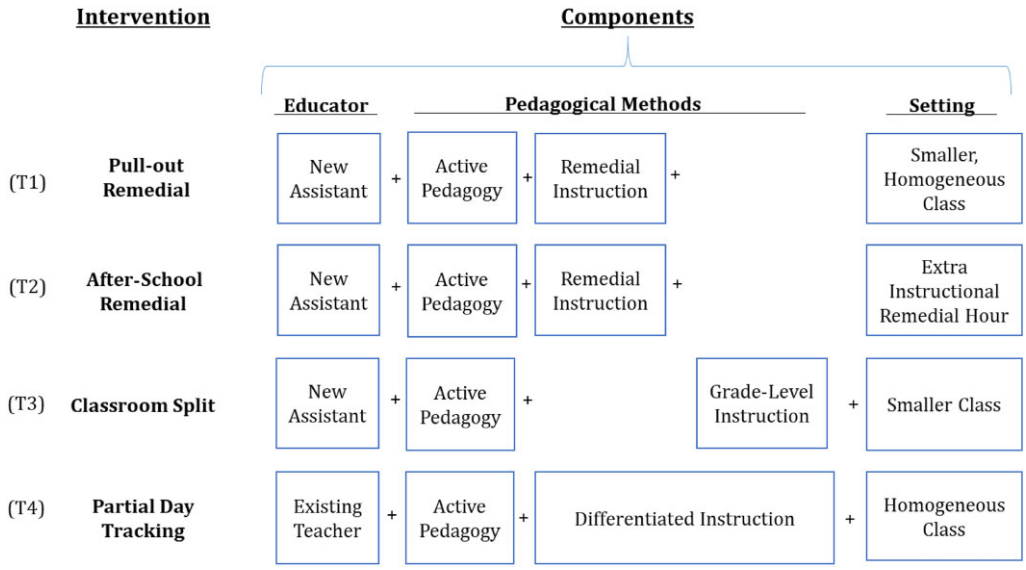

Fig. 2. *Intervention Components and Graphical Conceptual Framework.*

2.1.2. *After-school remedial*

The training, materials and testing and levelling procedures were the same as those for the assistants in the pull-out remedial arm. Instead of providing instruction during the school day as pull-out lessons, assistants met with remedial students after school for one hour a day, four days per week, providing an extra instructional hour on the days in which the lessons met. Assistants provided their lessons in school classrooms.

2.1.3. *Classroom split*

Classroom split assistants worked with an arbitrary half of each class on the official, grade-level, curriculum, one hour per day, four days per week, usually Tuesday through Friday. On Mondays, when new material is introduced in Ghanaian classrooms, the assistants attended the class to see what was being covered that week. During the classroom split time, the assistants removed half the class from the classroom and the classroom teacher remained with the other half of the class. Teachers should not have introduced new material during this time. Assistants were encouraged to work with a different group of students each time. Assistants were supposed to randomly select students, but we heard from assistants that they picked them arbitrarily, for example selecting every other student based on where students were sitting or their order in the classroom roster.[8] As with the pull-out remedial, assistants were encouraged to use extra space as available or meet outside under a tree if necessary.

[8] Assistants could have started to implement unconscious tracking, but this is unlikely as tracking is rarely done in this context. Prior to this study, even in schools with multiple streams (i.e., sections) of a single grade, schools did not track students, instead arbitrarily placing students in streams, typically based on their date of enrolment and an interest in keeping the numbers of students in each stream approximately equal. Furthermore, in the classroom split intervention, students were not tested by the assistants. The assistants likely had a sense of students' learning levels based on their interactions with students, but they would not have had formal data on achievement.

2.1.4. *Partial day tracking*

Teachers in the partial day tracking intervention received materials that spanned remedial to grade level and corresponding training to allow them to differentiate instruction across the three learning levels. Teachers tested students at the start of each term to determine their learning levels, assigning each student to learning level 1, 2 or 3. Students could change learning levels based on these termly exams or at any time at their teacher's discretion. For the first two terms of the intervention, partial day tracking occurred within a given classroom—teachers divided students by learning level within their classrooms focusing on one learning level at a time, while the other students worked independently. Starting the third term of implementation, partial day tracking occurred across classrooms—teachers divided their students across grades by learning level with one teacher teaching each learning level.[9] Tracking time was to occur one hour per day, four days per week. Students in this intervention had a more homogeneous classroom environment for this part of the school day.

2.2. *Conceptual Framework*

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

Comparing the two assistant-led remedial interventions (T1 versus T2 in Figure 2) shows the relative merits of the same lessons delivered either during or after school. In the during-school version (T1), remedial lessons led to a smaller, more homogeneous class setting for both the remedial and non-remedial students, but it deprived remedial students of some grade-level instructional time. In the after-school version (T2), remedial students had the same smaller, more homogeneous lessons as an extra instructional hour and the non-remedial students did not have a smaller, more homogeneous classroom hour. Both of these interventions were designed to shift the left tail of the learning distribution to the right. The comparison of the two during-school assistant interventions (T1 versus T3) shows the marginal effect of remedial versus grade-level instruction. The relative magnitudes of the pull-out remedial and the partial day tracking interventions (T1 versus T4) show whether a classroom teacher can replicate the benefits of an assistant who was working with fewer remedial learners by having the teacher focus on a homogeneous group of learners at their learning level, whether remedial or grade level. When comparing the after-school remedial to the classroom split (T2 versus T3), the difference compares an additional remedial, smaller, homogeneous instructional hour only for remedial learners relative to a smaller class size for an hour during the normal school day for all learners. The after-school remedial relative to the partial day tracking (T2 versus T4) shows the relative merits of an extra instructional hour focused only on remedial students versus more homogeneous instruction during the normal school day for all students. The final comparison of the classroom split relative to the partial day tracking (T3 versus T4) shows the relative effect of a smaller class size that involved an assistant versus a more homogeneous learning environment led by a teacher. These last two interventions

[9] Schools typically assigned the grade 1 teacher to level 1, the grade 2 teacher to level 2 and the grade 3 teacher to level 3. The revised method was communicated through a refresher training. No additional motivational tools or compensation was provided to the teachers.

were designed to shift the entire test score distribution to the right, not only focusing on remedial learners.

3. Empirical Strategy

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

$$y_{is} = \alpha + \beta\, treatment_s + \mathbf{X}'_{is}\boldsymbol{\Gamma} + \varepsilon_{is}, \tag{1}$$

where y_{is} is outcome y for individual i in school s, $treatment_s$ is an indicator variable equal to one if school s was a treatment school with a single indicator for all treatments (the control group is the omitted category), $\mathbf{X}_{is}$ is a vector of individual level controls and ε_{is} is a cluster-robust error term assumed to be uncorrelated between schools, but allowed to be correlated within a school. We always include dummy variables for strata (region by above/below-median pupil teacher ratio by above/below-median baseline test score) and gender in $\mathbf{X}_{is}$. When the outcome of interest is a student's test score, we implement a lagged dependent variable model and include the test score from the baseline as a control in the $\mathbf{X}_{is}$ vector.[10]

We additionally estimate the effect of each treatment,

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

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

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

Because of imperfect fidelity of implementation, we also perform an instrumental variable analysis. As with the non-instrumented version, we first consider an overall effect size, pooling all interventions into a single $treatment_s$ indicator, as we did in (1). In the first stage of this overall estimation, assignment to treatment at the school level is the instrument for whether we observed correctly formed groups during the spot checks,

$$groups_{is} = \eta + \gamma\, treatment_s + \mathbf{X}'_{is}\boldsymbol{\Phi} + \nu_{is}, \tag{3}$$

where $groups_{is}$ is the portion of spot checks in which school s attended by student i was implementing group learning, with the other notation as above. We then use the predicted value of $groups_{is}$ from (3) in the second stage

$$y_{is} = \alpha + \beta\, groups_{is} + \mathbf{X}'_{is}\boldsymbol{\Gamma} + \varepsilon_{is}, \tag{4}$$

with the notation as above. The coefficient of interest is the estimate of β, the estimated effect of the intervention on the test score of a student if their school was implementing group learning at all spot checks, i.e., $groups_{is} = 1$.

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

We also estimate the instrumented effect of each treatment, the analogue to (2), using the four exogenous treatment statuses to separately predict whether each type of group was meeting, effectively estimating the following equation four times, once for each type of grouping:

$$groups_{Tis} = \eta + \sum_{T=1}^{4} \gamma_T \, treatment_{Ts} + \mathbf{X}'_{is}\mathbf{\Phi} + \nu_{is}. \tag{5}$$

Here $groups_{Tis}$ is the portion of time that treatment T groupings, i.e., pull-out remedial, after-school remedial, classroom split or partial day tracking, were observed during spot checks at school s attended by student i, with the other notation as above. The second stage is then

$$y_{is} = \alpha + \sum_{T=1}^{4} \beta_T \, groups_{Tis} + \mathbf{X}'_{is}\mathbf{\Gamma} + \varepsilon_{is}, \tag{6}$$

with the notation as above. In this specification, the coefficients of interest are the estimates of β_T, the estimated effect of each intervention on the test score of a student whose school was observed grouping students according to treatment T at all spot checks. We did not observe any schools implementing any groupings other than the one to which they were assigned.

4. Sample Selection and Data

The 500-school experimental sample was nationwide in scope, including schools from all ten regions and 42 districts in Ghana.[11] From this sample, 100 schools were randomly allocated into each of the five treatment designations (four treatment arms and a control arm), stratified by region, above/below-median average baseline student test score and above/below-median pupil teacher ratio.

To evaluate the effect of the four interventions, we collected nine rounds of data across three academic years: a baseline, six spot checks and two achievement follow-ups. In the baseline and achievement follow-ups we administered surveys to head teachers (i.e., principals), teachers and students and tested students using bespoke exams in all 500 schools. The baseline occurred near the start of academic year 1 (October 2010), the first achievement follow-up was in academic year 2 (November 2011) and the second achievement follow-up was near the end of academic year 3 (July 2013). In academic year 1, we randomly sampled 25 students from grades 1 and 2 from those present on the day of initial enumeration. We attempted to follow these students through academic year 3 when they should have been in grades 3 and 4 if they progressed on pace.[12] The six spot-check rounds occurred termly, starting with the third term of academic year 1 (June 2010) and ending with the second term of academic year 3 (April 2013). In each of these data collection rounds, we visited a random subsample of schools, re-sampling with

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

[12] Students were encouraged to come to school on the days of the achievement follow-ups. Enumerators attempted to follow-up with all students who were absent, even those who had moved or were attending another school. We use data from an additional cohort of grade 1 students added at the first follow-up to calculate test score differences between grades. The test scores for these new students are statistically indistinguishable across treatments arms.

each round.[13] Data for these rounds recorded whether the school was implementing the intended intervention, assistant demographics, classroom activities, whether the student was still attending the particular school and in the expected grade, and student, teacher and head teacher attendance. For our instrumental variable strategy, we calculate the portion of spot-check visits in which we observed a school grouping its students in each of the four intervention methods. Figure 2 shows the data collection timeline. Online Appendix A.1 contains additional details on data collection and test design.

Data from our five treatment arms are balanced on student, teacher, school and assistant characteristics (see additional details in Online Appendix A.2). To provide some context, almost all students had shoes, but only one-quarter had a clean, good quality uniform. Even though the official age of entry for grade 1 is 6, grade 1 students were on average 7.8 years old near the start of the academic year. Attesting to the expansion of primary school access, about half of the sample had a literate father and about a third had a literate mother. Teachers were about 36 years old and about half were women. On average each grade had 37 students and one teacher. About one-quarter of schools had electricity. Assistants were about 25 years old and 40% were women. Almost all had completed high school and about one-third aspired to teach in the future.

Baseline achievement levels were low and heterogeneous within schools. At baseline, only about one-half of grade 1 students could correctly name a presented English letter and one-third could perform simple one-digit addition. At baseline, the average SD within grade 1 in a school was almost 90% of the average score difference between grades 1 and 2.

5. Results

5.1. *Student Outcomes*

5.1.1. *Achievement, selection into the test and persistence*

Table 1 contains the effects of the four treatments on the combined math and English student test scores based on (1) with a student's test score as the outcome. The sample is students who were grade 1 in academic year 1 when the treatment started.[14] Panel A combines all interventions into a single treatment indicator. Panel B contains separate estimates for each of the four interventions.

Relative to the control group, the treatments increased test scores by 0.08 SDs in the academic year 2 follow-up, after only about two terms of treatment (panel A, column (1)). Most of this gain was the result of the two remedial interventions that separately increased test scores by a statistically significant 0.11 SDs (panel B, column (1)). The other two interventions increased test scores by a positive, but statistically insignificant amount. We fail to reject the hypothesis that all the interventions had the same test score effect. The overall effect across all interventions is about 12% of a grade level of learning.[15]

As part of the motivation of the interventions was to improve foundational literacy and numeracy, in column (2) we restrict the exam to foundational literacy and numeracy questions.[16] The interventions on average increased foundational test scores by 0.11 SDs. For these foundational

[13] Each school was visited at least once with most schools visited three or four times.

[14] Our sample should have been in grade 2 in academic year 2 and grade 3 in academic year 3. We attempted to interview and assess all baseline students regardless of their grade at follow-up.

[15] At this follow-up students in the control schools in grade 2 had test scores that were 0.66 SDs higher than grade 1 students.

[16] These are the questions most similar to the Annual Status of Education Report (ASER) exam conducted in South Asia that has been used to evaluate similar interventions in India. 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.

Table 1. *Effects on Achievement in Math and English.*

	Academic year 2		Academic year 3	
	All questions (1)	Foundational questions (2)	All questions (3)	Foundational questions (4)
Panel A: interventions combined				
Any intervention	0.080**	0.109***	0.113***	0.127***
	(0.034)	(0.033)	(0.035)	(0.035)
Observations	8,654	8,654	8,004	8,004
R^2	0.54	0.45	0.47	0.38
Panel B: interventions separately				
(*1*) Pull-out remedial	0.106**	0.130***	0.143***	0.150***
	(0.043)	(0.042)	(0.047)	(0.048)
(*2*) After-school remedial	0.110**	0.151***	0.150***	0.162***
	(0.046)	(0.042)	(0.046)	(0.046)
(*3*) Classroom split	0.047	0.058	0.082*	0.073*
	(0.045)	(0.040)	(0.044)	(0.044)
(*4*) Partial day tracking	0.059	0.100**	0.077*	0.127***
	(0.046)	(0.044)	(0.046)	(0.044)
p-value of the test of equality				
1 = 2	0.93	0.65	0.88	0.80
1 = 3	0.20	0.08	0.20	0.12
1 = 4	0.32	0.49	0.18	0.64
2 = 3	0.20	0.03	0.15	0.06
2 = 4	0.30	0.27	0.13	0.45
3 = 4	0.80	0.34	0.92	0.23
Observations	8,654	8,654	8,004	8,004
R^2	0.55	0.45	0.47	0.39
Test score difference between grades	0.66	0.57	0.42	0.38

Notes: * Significant at 10%; ** significant at 5%; *** significant at 1%. Outcomes are SD test score changes. SEs clustered at the school level are reported in parentheses. All regressions include baseline test scores and dummy variables for strata and female. Sample of students in grade 1 tested at baseline (academic year 1) and relevant follow-up round. Difference between grades calculated based on control group means between grades 1 and 2 (academic year 2) and grades 2 and 3 (academic year 3). Columns (2) and (4): questions most similar to those appearing in the ASER. See the text for more details.

questions, the test score increases between the separate interventions are statistically significantly different: the two remedial interventions have larger, and statistically different, point values than the classroom split.

In academic year 3, students should have been in grade 3 and have received the intervention for two full academic years. When considered together, the interventions increased learning 0.11 SDs in academic year 3 (panel A, column (3)). As with academic year 2, the largest point values are for the remedial interventions (about 0.14 SDs; panel B, column (3)). Unlike in academic year 2, both the classroom split and partial day tracking increased achievement by a statistically significant amount (0.08 SDs). When limiting the analysis to the foundational questions, the point values are larger, and we reject the equality of the after-school remedial intervention relative to

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.

the classroom split (column (4)).[17] The effect size of 0.11 SDs is about 27% of a grade level of learning, over twice the grade-level-adjusted learning from academic year 2.[18]

One concern with any randomised controlled trial is that the effects are artificially generated by differential selection into test taking based on treatment status. We found no differential selection by treatment status or the interaction of treatment status and baseline achievement (Online Appendix Table A7). Nevertheless, we provide Lee (2009) bounds in Online Appendix Table A8, finding similar results as those in Table 1.

In Table 2, we show that the effects of the intervention persisted for students one year after they left the program. These students were near the end of grade 2 at the start of the program, received the intervention for all of grade 3 and should have been in grade 4 in the academic year 3 follow-up, and thus one year removed from the program. When considering all questions, including grade-level content, the pooled effect size is 0.07 SDs (column (1)) with specific effect sizes ranging from (statistically insignificant) 0.01 SDs (partial day tracking) to (statistically significant) 0.12 SDs (classroom split). We reject that the classroom split and partial day tracking coefficients are statistically equivalent. The focus in the classroom split intervention on grade-level content could have prepared students better for grade 4 content than the tracking or remedial interventions. The next column limits the exam to questions from the grade 1 through 3 curriculum. Receiving any treatment increased scores on average 0.10 SDs with point estimates for the specific treatment from (statistically insignificant) 0.05 SDs (partial day tracking) to (statistically significant) 0.13 SDs (classroom split) (column 2). We reject the claim that the classroom split and partial day tracking had the same effect at the 10% level. Column (3) further limits the questions to foundational content. The persistent effects on foundational material (column (3)) are similar to the foundational gains for students one year removed from the *balsakhi* program in Banerjee *et al.* (2007) even though the immediate effects after two years of these interventions were about half the size of the *balsakhi* program.

The benefits from the assistant interventions might have persisted, while those of the partial day tracking faded for at least three reasons. First, students received a higher effective dosage of the assistant interventions than the teacher intervention (see the additional discussion in Section 5.2 below), potentially resulting in stronger retention of material. Second, students could have changed their attitudes about school and learning by being exposed to an assistant who was about ten years younger on average than the teachers, resulting in more effective effort in the subsequent year. Third, the assistants were still present in the schools the year after the program, but not working directly with these students, potentially leading to additional encouraging interactions.

[17] Online Appendix Table A4 contains subject-specific test score effects for both the entire test (panel A) and the foundational content only (panel B). Online Appendix Table A5 repeats estimations of columns (1) and (2) of Table 1 with the cohort who started the program when they were in grade 2 and should have been grade 3 at the first follow-up. In all cases the point values are smaller for this cohort than those who started the program in grade 1. Online Appendix Table A6 combines the interventions in alternative ways—panel A considers differentiated (remedial and tracking) versus non-differentiated instruction (classroom split) and panel B divides the interventions by assistants versus teachers. In all cases the effect sizes relative to the control group are positive and statistically significant. For foundational questions, the differentiated instruction interventions had a statistically different and larger effect size than the non-differentiated intervention (panel A, column 2).

[18] At this follow-up, students in control schools who were in grade 3 had test scores that were 0.42 SDs higher than students in grade 2.

Table 2. *Persistent Achievement Effects in Math and English.*

	All questions (1)	Grade 1–3 questions (2)	Foundational questions (3)
Panel A: interventions combined			
Any intervention	0.074**	0.102***	0.104***
	(0.036)	(0.034)	(0.035)
Observations	4,302	4,302	4,302
R^2	0.49	0.43	0.42
Panel B: interventions separately			
(*1*) Pull-out remedial	0.072	0.102**	0.110**
	(0.045)	(0.043)	(0.044)
(*2*) After-school remedial	0.086*	0.114***	0.118***
	(0.046)	(0.044)	(0.045)
(*3*) Classroom split	0.120**	0.134***	0.133***
	(0.049)	(0.046)	(0.048)
(*4*) Partial day tracking	0.014	0.052	0.053
	(0.049)	(0.046)	(0.047)
p-value of the test of equality			
1 = 2	0.77	0.79	0.86
1 = 3	0.33	0.49	0.64
1 = 4	0.22	0.27	0.23
2 = 3	0.50	0.68	0.76
2 = 4	0.15	0.19	0.18
3 = 4	0.04	0.10	0.12
Observations	4,302	4,302	4,302
R^2	0.49	0.43	0.42
Test score difference between grades	0.45	0.47	0.51

Notes: * Significant at 10%; ** significant at 5%; *** significant at 1%. Outcomes are SD test score changes. SEs clustered at the school level are reported in parentheses. All regressions include baseline test scores and dummy variables for strata and female. Sample: grade 2 students tested at baseline (academic year 1) and in academic year 3. Students progressing on pace stopped receiving the intervention at the end of academic year 2. The test was invigilated at the end of academic year 3.

5.1.2. *Non-cognitive outcomes*

Based on the data from our unannounced spot checks, we test for the effect of the intervention on three non-cognitive outcomes: absenteeism, no longer attending school and grade repetition (Jackson, 2018). In control schools, about 36% of baseline students were absent, 24% were reported as no longer attending that school and of those who were still attending the school, 23% were in a grade below their expected grade.[19] The interventions did not change these non-cognitive outcomes. The full point values appear in Online Appendix Table A9.

5.1.3. *Heterogeneity by baseline characteristics*

The analysis thus far focused on the test scores of all students. Pull-out remedial, partial day tracking and classroom split changed instruction for all students either directly or as a result of some students being removed from a classroom. Even after-school remedial could have benefited the non-remedial learners by bringing their classmates closer to grade level. We test for

[19] For each student, we calculated the average portion of days absent across all spot checks, whether the school ever reported that the student was no longer attending that school (the sum of dropping out and transferring), or whether at any spot check the student was in a grade below what would be expected based on timely progression.

heterogeneity by both linear baseline test score and an indicator for likely non-remedial status.[20] The heterogeneity analyses in Online Appendix Table A10 and Online Appendix Figure A1 find a largely uniform effect.[21] One reason why we might find homogeneous effects across the baseline score distribution is that even the top students had limited literacy and numeracy at baseline. At the baseline, 54% of grade 1 students could read an upper or lower case English letter, 7% could read a three-letter word, 76% could recognise a one-digit number and 37% could do one-digit addition.

The intervention was not designed to favour one gender, yet gender might be a salient concern in a country with a gender bias in the assessment of teachers by head teachers (Beg *et al.*, 2021). Based on a simple comparison of means for the analysis sample, boys' test scores were 0.06 SDs higher than girls' scores at baseline. At the year 3 follow-up in control schools, this difference had widened—boys' scores averaged 0.10 SDs higher than girls' scores. In contrast in the treatment schools, boys' test scores were 0.03 SDs *lower* than girls' scores. In column 3 of Online Appendix Table A10 we formally test for heterogeneity in effects on the academic year 3 test scores by student gender by interacting the treatment variables with female. When using a single treatment indicator, the effect on boys' test scores is 0.07 SDs with girls' test scores increasing by an additional statistically significant 0.10 SDs (Online Appendix Table A10, panel A, column 3). For the three interventions with a remedial or tracking component, girls' test scores increased statistically more than boys' test scores by about 0.10 SDs for the pull-out and after-school remedial and 0.15 SDs for partial day tracking. Girls appear to have benefited more from homogeneous classrooms than boys, perhaps because girls were more hesitant to speak up to ask or answer questions in a heterogeneous learning environment. As teachers were 10 percentage points more likely than the assistants to be women, the additional improvement for girls in the assistant-led arms is not likely due to gender-matching role model effects.[22]

5.2. *Implementation and Treatment on the Treated*

The estimates in Section 5.1 were intention-to-treat estimates. Not all schools implemented the groups as intended or implemented them consistently, likely scenarios for other government implemented programs. We conducted six spot checks during the program implementation, recording whether groups were occurring and the type of student grouping schools were implementing. For each school, we use these data to calculate the portion of the visits that we observed a school correctly grouping its students. In this subsection, we measure the extent to which

[20] Across the spot checks, remedial intervention schools reported that on average 36% of the analysis sample students were in the remedial intervention. The other interventions were not asked about the remedial status of their students as this could have primed educators to think about providing remedial lessons. We approximate non-remedial status for all intervention arms as whether at baseline a student was in the top 64% of grade 1 students in his school.

[21] Within each intervention we do not find statistically distinguishable heterogeneity by baseline test score or remedial status (Online Appendix Table A10, panels A and B, columns 1 and 2). Relative to control schools, pull-out and after-school remedial statistically significantly increased test scores for both remedial and non-remedial students. In contrast, the classroom split only statistically significantly increased test scores for the non-remedial students and the partial day tracking only for the remedial students (panel B, column 2). Online Appendix Figure A1 shows the non-parametric distributional effects of the interventions combined (subfigure a) and then each intervention separately relative to the control group (subfigures b through e). The effects of the interventions were mostly positive for students across the baseline test score distribution.

[22] We also test for heterogeneity by whether the school had an above- or below-median average test score or above- or below-median pupil teacher ratio at the baseline and find no statistically significant heterogeneous effects (Online Appendix Table A11).

Table 3. *Fidelity of Implementation and Instrumental Variable Effects.*

		Instrumental variable estimates			
		Academic year 2		Academic year 3	
	Target educator teaching to a group (1)	All questions (2)	Foundational questions (3)	All questions (4)	Foundational questions (5)
Panel A: interventions combined					
Any intervention	0.257***	0.311**	0.421***	0.430***	0.485***
	(0.016)	(0.133)	(0.127)	(0.133)	(0.133)
Observations	500	8,654	8,654	8,004	8,004
R^2	0.22	0.55	0.45	0.47	0.38
Panel B: interventions separately					
(*1*) Pull-out remedial	0.315***	0.334**	0.409***	0.453***	0.468***
	(0.026)	(0.136)	(0.134)	(0.147)	(0.151)
(*2*) After-school remedial	0.401***	0.274**	0.375***	0.370***	0.398***
	(0.033)	(0.113)	(0.107)	(0.115)	(0.117)
(*3*) Classroom split	0.259***	0.176	0.216	0.300*	0.264*
	(0.027)	(0.171)	(0.156)	(0.161)	(0.160)
(*4*) Partial day tracking	0.056***	1.077	1.796**	1.284	2.099**
	(0.015)	(0.838)	(0.860)	(0.794)	(0.830)
p-value of the test of equality					
1 = 2	0.04	0.65	0.79	0.56	0.64
1 = 3	0.11	0.33	0.20	0.35	0.21
1 = 4	0.00	0.35	0.09	0.27	0.04
2 = 3	0.00	0.54	0.26	0.64	0.37
2 = 4	0.00	0.31	0.08	0.23	0.03
3 = 4	0.00	0.25	0.05	0.18	0.02
Observations	500	8,654	8,654	8,004	8,004
R^2	0.39	0.54	0.44	0.46	0.38
Control group mean or test score difference between grades	0.00	0.66	0.57	0.42	0.38

Notes: * Significant at 10%; ** significant at 5%; *** significant at 1%. SEs clustered at the school level are reported in parentheses. Additional controls: strata. Column (1): at the school level. Dependent variable defined as the portion of spot checks in which the target educator was teaching to the intended group. Columns (2)−(5): additional controls included the baseline test score and female. Instrumental variable estimates with treatment assignment at the school level as an instrument for groups meeting. See the caveats in the text regarding the estimates for partial day tracking.

implementation occurred and then use assignment to treatment as an instrument for a school implementing the program in a two-stage least-squares estimation.

The emphasis of the interventions was grouping students, whether by remedial status, by learning level or to have a smaller class size. In column (1) of Table 3 we test schools' fidelity of implementation of group learning. Each spot check overlapped with the time in which group learning should have been occurring in treatment schools. For each school, the portion of visits that the school was observed implementing group learning is the dependent variable. Overall, being in a treatment group increased the likelihood of students being divided into groups by 26 percentage points (Table 3, panel A, column (1)). The level of implementation was statistically different for each intervention, ranging from 6% of the time (partial day tracking) to 40% (after-school remedial), with the other two interventions in between (pull-out remedial at 32% and classroom split at 26%). We did not observe any control schools grouping their students. Below we provide evidence that the partial day tracking intervention increased student learning despite

infrequent implementation of grouping likely because the intervention improved other aspects of the learning environment.

Since groups did not meet as frequently as prescribed, students did not get the dosage intended. In the remainder of Table 3 we implement a two-stage least-squares strategy, as described in (3)–(6), with treatment status at the school level as an instrument for receiving the treatment, defined as the portion of spot-check visits in which students were correctly grouped. As with the previous results, we first combine all interventions into a single treatment indicator. These instrumental variable results should be interpreted with caution as they rely on three conditions that might not hold for all interventions. First, they assume linearity and scale in dosage, e.g., a 10-percentage-point increase in group learning is the same regardless of the base and doubling the portion of times that group learning was observed should double the effect sizes. Second, the exclusion restriction requires that the interventions only affect test scores through group learning. Third, implementing the group learning is a requirement for a strong first stage. We show below that these conditions are likely satisfied for the three assistant interventions, but that the partial day tracking results should be interpreted with caution.

Based on the instrumental variable (IV) estimates, if a school was observed implementing the program at all spot-check visits then student overall test scores would have increased 0.31 SDs in academic year 2 (panel A, column (2)) and 0.43 SDs in academic year 3 (column (4)). As with the non-instrumented version, the point values are larger when considering the foundational questions (0.42 SDs in academic year 2 and 0.49 SDs in academic year 3; columns (3) and (5)). Each of these estimates is about 3.8 times the size of the non-IV estimates in Table 1 and range from 48% to 128% of a year of learning. Panel B contains the coefficient estimates for each intervention individually, using the four school treatment status assignment indicators as instruments for the four separate types of groupings. Across both years and test levels, the remedial interventions increased test scores (columns (2) through (5)). Four other statistically significant findings appear in the IV estimates: the classroom split increased both test scores in academic year 3 (columns (4) and (5)) and partial day tracking increased foundational test scores in both years (columns (3) and (5)). The large point values on partial day tracking should be interpreted with great caution as the first stage is weak for this intervention and the instrument might violate the exclusion restriction (see further discussion below).

The IV estimates of the effects of the assistant-led interventions are robust to specifications designed to test the assumptions about linearity and scale in dosage and the exclusion restriction. First, to address concerns about linearity and scale in dosage, in Online Appendix Table A12 we redefine whether group learning was occurring as an indicator variable that takes the value of 1 if we observed a school implementing group learning in at least 50% of observations and 0 otherwise. About 26% of treatment schools were observed grouping their students at least half the time. When using this binary measure of implementation, the partial day tracking intervention no longer has a statistically significant effect on achievement, nor can we reject the hypothesis that its effects are equal to the other interventions. The other IV results are robust to this modification. Second, to address concerns about the exclusion restriction, we first show in Online Appendix Table A13 that the interventions at most minimally changed the operations of schools and classrooms outside of group learning, with the exception of those schools assigned to partial day tracking. In schools assigned to partial day tracking, the head teacher was more likely to be present and teachers were more likely to be in the classroom, engaged with students and using materials. Therefore, students in this intervention appear to have received more effective teaching even when group learning was not happening, leading to

the positive effects in Table 1 even though implementation of group learning was low in this arm.[23] Thus, the exclusion restriction is likely satisfied for the assistant interventions, but less assured for the partial day tracking. Online Appendix Table A14 provides revised IV estimates removing all schools assigned to the partial day tracking intervention. The IV results from Table 3 are robust to this change in sample. One final concern could be measurement error in the portion of time that group learning occurred as we only observed schools occasionally and not daily. Any measurement error should be unrelated to treatment status as the protocol for visiting each school was the same.

6. Cost Effectiveness

We base the cost effectiveness on the costs of the program using the ingredient method as the program was designed.[24] We provide alternative cost estimates based on the scale of the program and whether the assistant salaries are included.[25] We also provide alternative cost-effectiveness estimates, one scaled to the test score changes (Kremer *et al.*, 2013) and another to the years of schooling (Angrist *et al.*, 2020) per $100 scale based on the point estimates in Table 1, column (3). All costs are in 2019 US dollars.

First, considering the program at the implemented scale including the assistant salaries since the assistants were not used in the control schools, the per student costs over the span of the intervention were $161 for the remedial assistant interventions, $158 for the classroom split and $126 for the partial day tracking. The test score effects per $100 were 0.09 SDs for the pull-out remedial and after-school remedial, 0.05 SDs for the classroom split and 0.06 SDs for the partial day tracking. The years of learning effects per $100 were 0.21 years for the pull-out remedial, 0.22 for the after-school remedial, 0.12 years for the classroom split and 0.15 years for the partial day tracking.

Second, as the assistants were part of an existing government scheme, we provide cost estimates removing the assistant salaries. If assistant salaries are not included then the per student costs were $132 for the remedial assistant interventions and $128 for the classroom split. The costs and cost effectiveness for the partial day tracking are the same as above. The test score effects per $100 were 0.11 SDs for the remedial assistants and 0.06 SDs for the classroom split. The years of learning effects per $100 were 0.26 years for pull-out remedial, 0.27 years for after-school remedial and 0.15 years for the classroom split.

Third, at scale including assistant salaries would have per student costs of $39 for the remedial interventions, $38 for the classroom split and $21 for the partial day tracking. The test score effects per $100 were 0.36 SDs for pull-out remedial or partial day tracking, 0.22 SDs for the classroom split and 0.39 SDs for after-school remedial. The years of schooling per $100 were 0.87 years for pull-out remedial, 0.92 years for after-school remedial, 0.52 years for the classroom split and 0.86 years for partial day tracking.

[23] Given the low level of implementation of the grouping in the partial day tracking arm, its coefficients in Table 1 are the closest approximation to the effect of active pedagogy training alone. Because grouping was partially implemented, it is an upper bound.

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

[25] Teacher salaries were higher than assistant salaries. Per student costs decrease substantially when considering the per year cost of the program at scale as the fixed costs are spread among many more participants and materials have a multiple-year life span. We assume schools the same size as those in the study.

Fourth, at scale without the assistant salaries would have per student costs of $27 for the remedial interventions and $26 for the classroom split. The test score effects per $100 were 0.52 SDs for pull out remedial, 0.55 SDs for after-school remedial and 0.32 SDs for the classroom split. The years of schooling effects per $100 at this scale without assistant salaries would be 1.24 years for pull-out remedial, 1.31 years for after-school remedial and 0.76 years for the classroom split.

Regardless of the inclusion of the assistant salaries, the classroom split intervention is the least cost effective of the four interventions. If the assistant salaries are included then the similarity of the cost effectiveness at scale of the three interventions with levelled instruction is remarkable—the assistant-led remedial interventions cost approximately twice as much per student as the partial day tracking with approximately twice the benefit. If the assistant salaries are removed from the cost effectiveness calculation then the assistant-led remedial interventions were more cost effective than the teacher-led partial day tracking.

Because we have a multiple-year intervention considering the cost effectiveness of a shorter duration of the program is tempting. The first follow-up occurred near the start of the second school year, about two terms into the program. The effect size for the overall score at this first follow-up was about 71% of the point values for the second follow-up and the costs were less than half. Based on this metric, a shorter duration of the program could be considered more cost effective. If instead one considers the percent of a year of learning then the point value in academic year 3 is 27% of a year of learning, over twice the years of learning of 12% in academic year 2.

7. Discussion

7.1. *Implementation Lessons and Scale-Up*

Because of our extensive data collection over three academic years, this study contributes to the understanding of potential pitfalls, challenges and successes when implementing something entirely within existing government systems. One overarching lesson is that governments should invest in programs that have a high likelihood of implementation given existing constraints and realities. None of the interventions were implemented perfectly, yet the assistant interventions were implemented more than the teacher intervention. In a reticent civil service setting, as this was, assistants might be more effective. Alternatively, governments can think about ways to increase the likelihood that teachers implement the program with the same fidelity of implementation as the assistants. If governments could increase implementation then the effect sizes from using existing teachers could surpass that of the assistant arms.

7.1.1. *Common challenges and successes across all interventions*

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

interventions and had a strong focus on completing the annual curriculum. Weak oversight and delayed materials were two ways in which having this program implemented by the government likely muted the effect sizes across all four interventions.

7.1.2. *Unique challenges and successes for each intervention*

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

First, differences between the teacher and the assistant interventions likely contributed to different levels of implementation. Teachers had a dual mandate of teaching learning levels during partial day tracking and teaching grade-level material in regular curriculum lessons. Assistants only had the singular mandate to teach their group lessons. When teachers should have been engaging in partial day tracking, we most often observed them teaching curriculum lessons to their regular classrooms. Teachers might have reverted to the status quo because they felt pressure to complete the curriculum, were not being supported by their head teachers or were unhappy teaching only a single level of students. In contrast, assistants were hired specifically to teach the group lessons. Assistants were further on annual contracts, unlike the teachers' permanent contracts, which could have created stronger incentives to adhere to the intervention. Teachers were, however, more likely to be paid on time. Assistant salaries were delayed across all assistant interventions, especially in years 2 and 3. Taken together, teachers were the most likely educator to be present, but were the least likely to implement. The likelihood that the assistant interventions met as prescribed started at about 45% and decreased over time, while the likelihood that the teacher-led partial day tracking occurred was marginally higher in years 2 and 3 than year 1, but still less than 7% (Online Appendix Table A15).[26]

Second, the differences between the assistant interventions also contributed to differential implementation. The assistants who were the most similar to status quo teaching operations were more likely to be absorbed into those school operations instead of performing group teaching as intended. Classroom split assistants were the most similar to classroom teachers—teaching the same content as the teachers—and covered for a classroom teacher about 20% of the time they were present. This happened about half as often to the pull-out remedial assistants, who worked during school hours, but provided different content, and almost never happened to the after-school remedial assistants, who worked after school and provided different content. Overall, the classroom split intervention occurred less frequently than the other assistant interventions even though the classroom split assistants were the most likely to be present.

7.1.3. *Scaling the interventions*

In 2021, the Ghanaian Ministry of Education began scaling the partial day tracking intervention to 10,000 schools throughout the country as part of the Ghana Accountability for Learning Outcomes Project (GALOP), using this study and a subsequent study (Beg *et al.*, 2023) to support their decision. The Ministry was most interested in scaling an intervention that leveraged existing school-based staff and did not involve adding additional assistants. Beg *et al.* (2023)

[26] Given these trends in adherence, our estimates in Table 3 of adherence to group learning could be biased if schools were sampled differentially by treatment status and year. In each year, the likelihood of a visit was not differential by treatment status. In years 1 and 3 some schools have incomplete classroom observation data, marginally differential by treatment status. To avoid bias due to changing adherence over time, we repeated the IV estimates using only the adherence to grouping data from the year 2 spot checks as the endogenous variables. These results are very similar to the full sample estimates in Table 3 and appear in Online Appendix Table A16.

showed that involving school management structures in the partial day tracking intervention improved the fidelity of implementation and led to larger test score increases.

7.2. *Comparison to Other Studies*

Recall from Section 1 that this study combined many of the lessons of the last 20 years in improving student learning into a single RCT implemented within existing government systems. Based on the previous studies implemented within an existing government system (Bold *et al.*, 2018) or without an extra supervisory layer (Banerjee *et al.*, 2017), whether any of the four interventions when implemented by the government would increase student learning was unclear. Online Appendix Table A17 contains some of the common steps along the causal chain that are reported across studies—educators being trained, being present, using intervention materials and adhering to the intervention as measured by whether the correct learners were being taught with the prescribed method—between the previous implementations of class size reductions, full and partial day tracking, and pull out and after-school remedial instruction.[27] The educators in this study were more likely to be trained, but less likely to be present. Relative to previous implementations that improved test scores, they were less likely to be using intervention materials or adhering to the interventions.

Online Appendix Figure A2 plots our effect sizes (solid bars) relative to previous interventions that are the most similar to ours. Based on existing evidence, whether our four interventions could be effective without NGO support was unclear, as can be seen by the small and statistically insignificant effect sizes of the only prior government intervention (Bold *et al.*, 2018) and the two teacher-led differentiated instruction versions in India without an extra supervisory layer (Banerjee *et al.*, 2017). Furthermore, the assistant-led interventions had not previously been tested in an exclusively government implemented program. We show that existing systems can increase learning.

7.3. *Mechanisms and Policy Implications*

Using the conceptual framework and comparisons across the interventions from Section 2.2 and Figure 2, the four interventions provide six insights into the estimated mechanisms behind improvements in student learning with policy implications for future iterations of similar interventions. Unfortunately, the relative effect sizes of perfect implementation cannot be known, instead the realised implementation is what could be expected when implemented within similar existing government systems. In discussing mechanisms, we start with the educator (the left side of Figure 2) and end with the instructional setting (the right side of Figure 2) and an overarching lesson.

(*1*) The differences between the point values cannot be explained only by whether the intervention educator was an assistant or an existing teacher. The two most effective interventions were led by assistants (pull-out remedial and after-school remedial). The third assistant intervention (classroom split) had a very similar point value to the teacher-led partial day tracking.

[27] Blanks in the table denote that the study did not report this statistic while 'N/A' indicates that the particular study did not include this element (Banerjee *et al.*, 2016). For example, the tracking and classroom split interventions in Kenya did not include materials.

(*2*) The pedagogical method of remedial instruction, whether exclusively remedial or as partial day tracking, led to larger point values for the foundational content than the classroom split intervention that only focused on grade-level instruction.
(*3*) The smaller, more homogeneous hour for all students in pull-out remedial is almost an exact substitute for an extra instructional hour for remedial students. Both pull-out remedial and after-school remedial had the same instructional content, but different delivery mechanisms, with almost identical point value effects (0.008 SDs different).
(*4*) A smaller class size with remedial content had a larger point value effect on learning than a smaller class size without remedial content.
(*5*) While still statistically different from 0, the teacher-led partial day tracking intervention had the smallest effect size point values. It was also implemented about one-fifth as often as the other interventions. Had it been implemented to the same degree as the other interventions, the effect sizes might have exceeded the others.
(*6*) The assistant interventions conferred longer-lasting gains—all three had statistically significant effects for foundational and grade 1 through 3 content in the year after the intervention and the after-school remedial and classroom split let to statistically significant score increases for the whole test. The higher dosage that students received in these arms could have led to the longer-lasting gains. Furthermore, learning from another adult who was not the regular teacher and on average ten years closer in the age to the student could have inspired additional interest in learning. Finally, these assistants were still present at the school, but no longer working with these students, in the year after the program, leaving the potential for additional interactions and check-ins.

The overarching lesson is that all four interventions increased learning relative to the status quo. Therefore, for a government interested in increasing student learning, involving assistants in during- or after-school remedial instruction or as para-teachers reinforcing the regular curriculum or having existing teachers provide a more active classroom and partial day tracking are all viable options. The most effective intervention in a particular context likely depends on which intervention could be implemented with the highest fidelity.

8. Conclusions

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

Showing that governments can improve productivity in government primary schools, all four interventions increased student learning based on a test administered at the end of grade 3 for those students who started the program near the end of grade 1. The average effect of the treatments was 0.11 SDs. The interventions' positive effects persisted for those students exposed

to the program in grades 2 and 3 and tested at the end of grade 4, one year after ending the program. We find no evidence that the program affected the non-cognitive outcomes of student attendance, drop-out or likelihood of being demoted. Taking into account imperfect compliance by using a TOT estimate, a fully implemented program would have increased test scores an average of 0.4 SDs. When considering cost effectiveness from the intention-to-treat estimates, the after-school remedial, pull-out remedial and partial day tracking interventions were similarly cost effective—the effect sizes and costs of the first two were approximately twice the size of the third.

All interventions faced issues of material delays, teacher and assistant absenteeism, and weak mechanisms for support and monitoring, factors that could potentially be remedied with additional training and support for managerial layers of the civil service. Stronger adherence to the intervention as prescribed could result in larger effect sizes.

Innovations for Poverty Action, USA
Omidyar Network, USA
University of Delaware, CGD, J-PAL & NBER, USA

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Angrist, N., Evans, D.K., Filmer, D., Glennerster, R., Rogers, F.H. and Sabarwal, S. (2020). 'How to improve education outcomes most efficiently? A comparison of 150 interventions using the new learning-adjusted years of schooling metric', Policy Research Working paper, World Bank.

Angrist, N. and Meager, R. (2022). 'The role of implementation in generalisability: A synthesis of evidence on targeted educational instruction and a new randomised trial', Syntheses Working Paper 4, CEDIL.

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

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

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M. and Walton, M. (2016). 'From proof of concept to scalable policies: Challenges and solutions, with an application', Working Paper 22931, National Bureau of Economic Research.

Banerjee, A.V., Cole, S., Duflo, E. and Linden, L. (2007). 'Remedying education: Evidence from two randomized experiments in India', *The Quarterly Journal of Economics*, vol. 122(3), pp. 1235–64.

Bau, N. and Das, J. (2020). 'Teacher value added in a low-income country', *American Economic Journal: Economic Policy*, vol. 12(1), pp. 62–96.

Beg, S., Fitzpatrick, A. and Lucas, A.M. (2021). 'Gender bias in assessments of teacher performance', *AEA Papers and Proceedings*, vol. 111, pp. 190–5.

Beg, S., Fitzpatrick, A. and Lucas, A.M. (2023). 'Managing to learn', Working Paper 31757, National Bureau of Economic Research.

Beg, S.A., Lucas, A.M., Halim, W. and Saif, U. (2022). 'Engaging teachers with technology increased achievement, bypassing teachers did not', *American Economic Journal: Policy*, vol. 14, pp. 61–90.

Bold, T., Kimenyi, M., Mwabu, G., Ngángá, A. and Sandefur, J. (2018). 'Experimental evidence on scaling up education reforms in Kenya', *Journal of Public Economics*, vol. 168, pp. 1–20.

Duflo, A., Kiessel, J. and Lucas, A.M. (2023). 'Experimental evidence on four policies to increase learning at scale—replication package', https://doi.org/10.5281/zenodo.10492707.

Duflo, E., Dupas, P. and Kremer, M. (2011). 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *The American Economic Review*, vol. 101(5), pp. 1739–74.

Evans, D.K. and Acosta, A.M. (2021). 'Education in Africa: What are we learning?', *Journal of African Economies*, vol. 30(1), pp. 13–54.
Evans, D.K. and Yuan, F. (2022). 'What we learn about girls' education from interventions that do not focus on girls', *The World Bank Economic Review*, vol. 36(1), pp. 244–67.
Gilligan, D.O., Karachiwalla, N., Kasirye, I., Lucas, A.M. and Neal, D. (2022). 'Educator incentives and educational triage in rural primary schools', *Journal of Human Resources*, vol. 57(1), pp. 79–111.
Hartwell, A. (2010). 'National literacy acceleration program (NALAP) implementation study', Working paper, Education Quality for All Project (EQUALL).
Jackson, C.K. (2018). 'What do test scores miss? The importance of teacher effects on non–test score outcomes', *Journal of Political Economy*, vol. 126(5), pp. 2072–107.
Kremer, M., Brannen, C. and Glennerster, R. (2013). 'The challenge of education and learning in the developing world', *Science*, vol. 340(6130), pp. 297–300.
Krueger, A.B. (1999). 'Experimental estimates of education production functions', *The Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.
Lee, D.S. (2009). 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *The Review of Economic Studies*, vol. 76(3), pp. 1071–102.
Lucas, A.M. and Mbiti, I.M. (2012). 'Access, sorting, and achievement: The short-run effects of free primary education in Kenya', *American Economic Journal: Applied Economics*, vol. 4(4), pp. 226–53.
Ministry of Education. (2014). 'Ghana 2013 national education assessment', Technical Report, Ministry of Education.
Muralidharan, K., Niehaus, P. and Sukhtankar, S. (2016). 'Building state capacity: Evidence from biometric smartcards in India', *American Economic Review*, vol. 106(10), pp. 2895–929.
Muralidharan, K. and Niehaus, P. (2017). 'Experimentation at scale', *Journal of Economic Perspectives*, vol. 31(4), pp. 103–24.
Nickow, A., Oreopoulos, P. and Quan, V. (2020). 'The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence', Working Paper 27476, National Bureau of Economic Research.

The Economic Journal, **134** (*July*), 1985–2008 https://doi.org/10.1093/ej/ueae003 © The Author(s) 2024. Published by Oxford University Press on behalf of Royal Economic Society.
Advance Access Publication Date: 2 February 2024