Public Servants Under Pressure: Experimental Evidence on Efforts to Improve Teaching in India[*]

Andreas de Barros

Johanna Fajardo-Gonzalez

Paul Glewwe

Ashwini Sankar

January, 2021

We study a large program that seeks to improve mathematics learning in public primary schools in India. In a cluster-randomized trial, two treatment arms promoted activity-based instruction by providing teaching materials and teacher training. One of these arms also promoted community engagement through community-led student contests. A third arm remained untreated. After 13 months, the version without contests improved teaching quality and learning (predominantly among girls). Both versions improved student attitudes towards math. Yet, the addition of contests— which are intended to put pressure on teachers to increase their students' performance— worsened instructional quality (especially classroom culture), and we can rule out that the contests added even small improvements in learning. *JEL* Codes: C93, I21, I25.

[*] This study has been conditionally accepted at the Journal of Development Economics (JDE), through its pre-results review process. The authors thank the editor and two anonymous reviewers for helpful comments; the authors also thank the editor for his encouragement to submit the article to another journal. Abhijit Banerjee, Julián Cristia, Alejandro Ganimian, and Tavneet Suri provided additional comments. The study is pre-registered at the AEA RCT Trial Registry (AEARCTR-0003494). We gratefully acknowledge the funding provided by the Omidyar Network for this project. The study has been approved by the University of Minnesota Human Research Protection Program (Study 4101) and by the Institute for Financial Management and Research (IFMR) Human Subjects Committee. The authors thank J-PAL South Asia and its team of field staff. Sandhya Seetharaman, Prajwal Shenoy, and Anuja Venkatachalam provided excellent research assistance (Shenoy) and outstanding research management (Seetharaman and Venkatachalam). The authors thank staff at Akshara Foundation for their collaboration, in particular Ashok Kamath and K. Vaijayanti. The authors are grateful for the collaboration between Akshara Foundation and the Government of Karnataka. The usual disclaimers apply. The authors have no conflicting interests to declare.

I. Introduction

The education sector illustrates how, in many developing countries, public service delivery is plagued by low productivity. In recent decades, many developing countries have substantially increased their spending on education, which was followed by increased enrollment in primary education. Despite—or perhaps because of—these developments, student learning levels remain very low, and researchers have shifted their attention to the low academic performance of primary school students.

India exemplifies this phenomenon of increased education spending, high student enrollment rates, and low levels of student learning in public primary schools. Government spending on education in India more than doubled between 2006 and 2013 (in constant PPP$; see UNESCO Institute for Statistics 2018). Alongside this increased spending, India's primary school enrollment rates have consistently been over 95 percent for both boys and girls over the past decade (ASER 2018). Yet, only about half of Indian children enrolled in *grade five* can read a simple paragraph at the *second-grade* level (50.1 percent of children), or solve a two-digit subtraction problem (52.3 percent of children) (ASER 2018). These alarming statistics have opened a serious debate on "what works" to improve learning in India.

This paper investigates the causal effects of two approaches to increase student learning in public primary schools: (1) provision of teaching materials and teacher trainings designed to improve instructional quality, and (2) provision of teaching materials and teacher trainings *and* implementation of community-led student contests that attempt to raise parent and community engagement with their local schools. The first approach hypothesizes that poor public service delivery stems from (low) state capacity and human resource development. Providing public servants with additional resources, and training on how to use them, will increase their productivity (Migdal 1988; Armstrong 2000; Acemoglu, García-Jimeno, and Robinson 2015). The second contends that, for increases in capacity and human resources to improve public service delivery, and in particular to raise student learning, increased public pressure is needed in order to change the incentive structures of public servants, such as teachers. Multitasking models suggest that this strategy works only if the changed incentives are aligned with higher-productivity behaviors; the will be ineffective, or maybe even harmful, if they cause public servants to shift their effort to lower productivity tasks (Holmstrom and Milgrom 1991; Dixit 2002).

We evaluate a large, state-wide program in Karnataka, India. The program promotes activity-based instruction that aims to enable students to learn mathematical concepts and develop their mathematical thinking through engaging activities that allow them to find creative ways to solve mathematical problems—in marked contrast to conventional chalk-and-talk methods commonly used in Indian schools. The program also conducts community-led contests, where student test scores are made public, to increase community engagement. It is a collaboration between the state government and an Indian non-governmental organization (Akshara Foundation) that includes a phased scale-up to all 44,000 public primary schools in the state.

We implemented a cluster-randomized trial to estimate the causal effect of this program on student learning in mathematics. We assigned 98 administrative units (*Gram Panchayats*[1]) and their schools to either the program or a control group. To disentangle the effect of the community contests from the teacher training and teaching materials, we conducted a second randomization of treated Gram Panchayats. Our sample of 292 schools in two districts includes all students in grade four in those schools at the start of the study.

[1] The local government system in India, at the village or town level.

We begin by investigating adherence to treatment assignment and implementation fidelity. We find that: 1. All program schools received the additional teaching inputs; 2. Almost all (89 percent) of grade-four teachers in the program schools received the program's training; 3. After program implementation, there were large differences in the pedagogical methods used by program school teachers (relative to control group teachers); and 4. The vast majority (86 percent) of the program schools assigned to the community contest group participated in those contests. Any lack of program impact is thus unlikely to be due to failure to implement the program. We also find support for the study's internal validity, including experimental balance and (absence of) attrition bias.

We then present four sets of results. First, we show intention-to-treat (ITT) effects across both program variants. After 13 months of the program, we find that its combined effect had at most small, statistically insignificant impacts on fourth-grade students' learning of mathematics. Specifically, combining both program variants, we estimate an average impact of 0.07 standard deviations (SDs) of the distribution of test scores on students' mathematics skills.

Second, since this finding may mask differences in the two program variants, we report findings by treatment arm. We find that the variant without contests raised student learning by 0.12 SDs ($p < 0.1$). In contrast, the estimate for the variant with community contests is almost zero (0.02 SDs), and we can rule out that contests added sizeable learning increases over and above the variant with no contest (added effects of 0.07 SDs or more ruled out at 95 percent confidence).

Next, we examine in detail potential mechanisms. We use four rounds of classroom observations, one-on-one pupil interviews, and home visits to parents, to estimate impacts on three sets of intermediate outcomes: student attitudes on mathematics, instructional quality, and parent engagement. We find that both program versions raised student attitudes towards mathematics (0.09

SDs). Both also succeeded in promoting teaching practices that are expected to promote socioemotional skills (0.17 to 0.24 SDs). However, the community contest version created a more hostile classroom environment (-0.14 SDs, for the overall study period). Adding contests to the intervention led to sizeable negative effects on this dimension of teaching, in the study periods after the community contests had been conducted (-0.30 to -0.48 SDs). Parents and teacher interviews also suggest that the version with community contests did not increase parents' engagement in their children's education.

Fourth, analyses of heterogeneous effects by gender reveal a significant impact of 0.14 standard deviations for girls' math scores, but no effect for boys. These results are driven by the variant without contests, which raised girls' math scores by 0.18 standard deviations ($p < 0.05$). This finding is robust to several checks that account for attrition, alternative sample definitions, and an alternative approach to measuring the main outcome of interest. One robustness check (employing randomization inference) yields a p-value somewhat above the 10-percent critical value ($p = 0.15$) for girls' math scores. Yet, our finding that the addition of community contests did not lead to higher impacts is robust to all robustness checks. We conclude by highlighting the cost-effectiveness of the program variant without community contests. The cost of this variant is about USD 7 per student (or about USD 14 if all costs are attributed to girls, given the null findings among boys). These costs are approximately one-tenth (one-fifth) of the per-student costs required for a successful intervention recently evaluated in Indian government schools in Rajasthan (Muralidharan and Singh 2019). More specifically, the Rajasthan program raised mathematics test scores by 0.21 SDs, at a cost of USD 66 per student per year. In comparison, if all costs from our program are attributed to girls, it increased their test scores by 0.18 SDs at a cost of USD 14.

This paper makes three contributions to the literature. The first is to the broader literature on state capacity and public-servant quality in developing countries (Pritchett, Woolcock, and Andrews 2013; Best, Hjort, and Szakonyi 2017; Bertrand et al. 2020). In particular, we complement other evaluations of large-scale interventions that seek to improve teacher effectiveness in public schools. Prior evidence suggests that teacher capacity building is more effective if it provides detailed guidance on what teachers should teach, and how to adjust their pedagogy (Popova, Arancibia, and Evans 2016; Conn 2017; Ganimian and Murnane 2016). Successful on-site training and teacher coaching programs also support this finding (Cilliers, Fleisch, Prinsloo, et al. 2020; Majerowicz and Montero 2018). Yet, it is not clear how to ensure the effectiveness of said programs at scale, including when intensive support is removed (Banerjee et al. 2017), responsibilities are transferred from non-governmental organizations (NGOs) to the government (Bold et al. 2018; Duflo, Kiessel, and Lucas 2020), and training is provided remotely (Cilliers, Fleisch, Kotzé, et al. 2020).

Our second contribution is to the literature on whether supply-side programs to raise public servants' productivity have complementarities with demand-side interventions to increase community engagement. The evidence on this question is mixed. In public health, Björkman and Svensson (2009) and Björkman Nyqvist, Walque, and Svensson (2017) find strong, lasting, positive effects of a Ugandan intervention that provided communities information on the quality of services at local government-run health centers. However, Raffler, Posner, and Parkerson (2020) document how these effects were not found for a similar intervention also implemented in Uganda. In education, several studies report positive effects on student learning from interventions promoting community and parental engagement in Bangladesh (Islam 2019), India (Pandey, Goyal, and Sundararaman 2011), Indonesia (Pradhan et al. 2014), Kenya (Duflo, Dupas, and Kremer 2015), Mexico (Gertler, Patrinos, and Rubio-Codina 2012), and Niger (Aker and Ksoll 2019). Yet, other community-focused programs did not increase child learning in the Gambia (Blimpo, Evans, and Lahire 2015), India (Banerjee et al. 2010), and Mexico (Barrera-Osorio et al. 2020). Finally, evidence from Ghana suggests that increasing parental involvement can reduce the effectiveness of education interventions if parents disagree with novel pedagogical practices (Wolf et al. 2019).[2]

Third, our study also connects to a nascent economics literature that evaluates at-scale public policy through randomized experiments (Muralidharan and Niehaus 2017; Duflo 2020). We evaluate the program as implemented on a very a large scale, in government schools, with public teachers, during the usual school hours. We thus add to research on the effectiveness of public programs under government leadership, going beyond smaller, tightly controlled pilots that could suffer from site selection bias (Allcott 2015), implementer effects (Vivalt 2020), and publication bias (DellaVigna and Linos 2020).

The rest of the paper proceeds as follows. Section II describes the experiment, its sampling, and the randomization. Section III presents the outcomes, hypotheses, measurement methods, and data used. Section IV explains the empirical strategy. Section V presents the results and Section VI concludes.

II. Research Design

II.A. Context

From ages 6 to 14, schooling in India is compulsory, free, and a fundamental right (Ministry of Law and Justice 2009). Elementary education runs from grades 1-8, with grades 1 to 5 referred to as "primary" and grades 6 to 8 referred

[2] Sexton (2020) also finds unintended effects of a community-focused intervention in Peru, including reduced local participatory budgeting.

to as “upper primary” education. In 2018, India’s school system had 1,255,841 schools serving “primary” grades, of which more than two thirds (69 percent, or 860,790) were managed by state and local governments (NIEPA 2018).

We conducted this study in partnership with the Akshara Foundation, a large NGO that is dedicated to ensuring quality pre-school and primary education in India. Founded in the year 2000, Akshara has agreements with several state governments to provide support to primary education in government-led schools.

We implemented this study in the Indian state of Karnataka, which is an ideal context to conduct a state-wide proof of concept for education interventions that may be scaled up to the entire country. First, the state is large, ranking sixth in terms of area and eighth in population (MHA 2012). Second, Karnataka exemplifies how increased enrollment and additional inputs may not coincide with improved student learning. It ranks near the top in terms of student enrollment (over 99 percent of rural children ages 5 to 14 are enrolled in school), attendance (the observed attendance of rural primary students and teachers is over 90 percent) and infrastructure (e.g., over 99 percent of rural primary schools have a library or dedicated reading corner (ASER 2018; NIEPA 2018). Yet, arithmetic performance of primary school students ranks Karnataka near the bottom of India’s states (e.g., less than 20 percent of rural government-school students in grade 5 can do basic division) (ASER 2018). Third, other states mimic at least some of Karnataka’s education policies; for example, Odisha recently adopted the intervention we evaluate in this paper.

II.B. Intervention

The Akshara Foundation’s Ganitha Kalika Andolana (GKA) intervention combines the provision of new instructional materials, related teacher training, and community engagement to improve primary-school students’ mathematics abilities. This subsection describes each of the program's two main components.

The program was started in 2011 for government primary schools in one block of Bangalore Rural District.[3] According to the Akshara Foundation, Karnataka's Government has since committed to scale up the program to all of the state's 44,000 Government primary schools, in a phased manner. Moreover, in 2017 another Indian state, Odisha, began to implement GKA, and had expanded it to about 30,000 schools by 2020.

II.B.1. Teaching inputs for activity-based instruction, and related training

The program's first component consists of providing additional teaching inputs, and related teacher training. This component seeks to refocus mathematics instruction on conceptual understanding, rather than rote learning. Specifically, GKA provides a kit of teaching-learning materials (TLMs), and instructions to teachers, to facilitate activity-based pedagogy.[4] The TLM kits include items such as an abacus, a series of shapes, and measuring kits. Each item maps into mathematical concepts that are required by the state curriculum.[5]

A pool of expert teachers provide training to the primary school teachers.[6] The training is designed to enable teachers to create activities using the TLM kit's items. In addition to this initial training, a field coordinator, appointed at the block level, supports the teachers as they implement this new teaching method.

[3] It started in 257 government primary schools in the Hoskote block, near Bangalore.

[4] This pedagogical approach follows a "concrete-representational-abstract" (CRA) model, where students are expected to, first, develop conceptual understanding by manipulating objects; thereafter, learn how pictures, numbers, and symbols represent objects; and finally, master mathematical problems using only abstract numbers and symbols. CRA is loosely based on a learning theory that has three "Stages of Representation": enactive, iconic, and symbolic learning of mathematics (Bruner and Kenney 1965).

[5] This mapping is "many-to-many": a concept may be learned from multiple TLMs and one TLM may teach multiple concepts.

[6] Off-site training sessions are held during the state's teacher training schedule, replacing its content. There are no separate training sessions, which keeps costs neutral.

II.B.2. Community contests

The program's second component is the community contests. These Gram Panchayat Mathematics Contests ("GP contests") convene stakeholders to witness the mathematical performance of school children, during a public assessment. Contests start with a math test for the community's students—they can be from any government primary school in the GP. Following the test, participants discuss the GKA program and related education issues (focusing on students' learning outcomes and the quality of instruction they receive). Next, the assessment results are announced, the top three students are recognized, and other education performance statistics are presented to community members. While the Akshara Foundation initiates these contests in participating GPs, the GP and other local sources pay for all operational expenses. In any given school year, a GP holds at most one contest.

II.C. Sampling and sample

II.C.1. Sampling

We implemented the study in two districts in Karnataka: Tumkur and Vijayapura.[7] We purposely selected these two districts to maximize the study's geographic spread and representativeness, within the state. In a first step, we randomly sampled 98 Gram Panchayats (GPs) from these two districts. Within each GP, we then randomly sampled three schools, for a total of 294 schools. Two schools were removed thereafter, reducing our sample to 292 schools, after baseline data collection revealed that they had no fourth-grade students.[8]

Prior to sampling, we used administrative data to exclude some schools and GPs. First, in order to track students into higher grades, we focus on

[7] In India, districts are the largest administrative units within a state or territory. Karnataka has 30 districts.

[8] These schools were removed prior to randomization into treatment and control schools.

one type of government primary school: “Higher Primary Schools” (HPs).[9] Second, we include only HPs with the following characteristics: (1) the medium of instruction is Kannada (87.5 percent of HPs); (2) the lowest grade is grade four or lower (99.9 percent); and (3) grade four had at least five students in the previous school year (88.9 percent). Finally, for logistical reasons, we include only GPs with at least three eligible schools (84.5 percent of eligible HPs).

The sampling strategy ensured that half of the study’s GPs and schools were drawn from each of the two districts. Beginning with a roster of all GPs in these districts, our first step was to randomly select 49 GPs from each district. This was done using “probability-proportional-to size” (PPS) sampling, where a GP’s selection probability reflects its number of eligible schools.[10]

The second step consisted of randomly selecting three schools from each of the 98 GPs. Within each GP, all schools had the same probability of being selected. Finally, we included all fourth-grade students in these sampled schools (as measured at baseline). Appendix Figure A1 depicts the study’s two districts and Appendix Figure A2 depicts its randomly selected GPs and schools.

II.C.2. Sample and sub-samples

Sample. Baseline data collection revealed that 5,227 fourth-grade students were formally enrolled in the study’s 292 schools. Of those, 4,026 (77.0 percent) were present during the baseline data collection.[11]

We consider these 4,026 students as the study’s sample. Our baseline data indicate that, on average, these students were about 9 years and 2 months old.

[9] Most (70.4 percent) HPs end with grade seven; about a quarter (25.1 percent) end with grade eight. About half (45.6 percent) of Karnataka’s government primary schools are HPs; the rest are “Lower Primary Schools” (LPs), which serve grades one to five.

[10] Given their large size, three GPs were included with certainty (two in Tumkur and one in Vijayapura). All other GPs were selected using PPS.

[11] This number is similar to other, large-scale, nationally representative assessments in India. For example, Goodnight and Bobde (2018) report a 73.1 percent attendance rate for India’s government primary schools.

About 53.0 percent of the sample is female.[12] Of these 4,026 students, 3,971 (98.6 percent) took the written baseline test, and 3,881 students (96.4 percent) took the study's written *and* oral baseline tests (we describe these tests further below, in Section III). Our analyses focus on the students with both tests, but we also show robustness checks for the sample without the oral baseline test, and for the full sample (including those without a written baseline test).

Sub-samples. To analyze intermediate outcomes, we conducted interviews with sub-samples of students and parents. This was done by randomly selecting (up to) eight students per school, using the baseline roster and students' performance on the baseline test.

More specifically, we stratified each school's list of students by: (a) gender (female/male): and (b) baseline performance (above/below the school median). We then randomly selected two students per stratum.[13] We used the same procedure to generate a separate sub-sample of parents; however, for logistical and budgetary reasons, we only selected four parents per school. We repeated these sub-sampling procedures separately for each survey round.

II.D. Randomization

II.D.1. Randomization of treatment and control units

To increase statistical power and ensure balance across treatment and control units, we conducted a stratified randomization to assign the 292 schools to be treatment or control schools. After the baseline test, within each district we

[12] These students' average age and gender are approximate numbers, since this information is missing for 2.0 percent of the students.

[13] More specifically, we selected all enrolled students if there were less than eight in a given school. We randomly drew more from the school's remaining strata if any given stratum had less than two students. For example, in an all-girls school, we randomly selected four girls with a baseline test score above the median, and four girls with a baseline test score below the median.

used baseline test scores to create quadruplets of GPs with similar academic performance.[14] Next, for each stratum of four GPs, two were randomly selected to participate in the GKA program, while the other two remained as "controls."[15] Thus, 49 GPs and their selected schools were assigned to receive the program; the remaining 49 and their selected schools continued with "business-as-usual."[16]

We repeated the above-mentioned randomization procedure ten times, to select the one with greatest balance. To do this, we selected a vector of covariates—from India's District Information System for Education (DISE)—that are predictive of baseline scores. We then calculated t-statistics for the difference of each of the selected variables across the two groups of GPs, as well as the baseline math score. We did so by regressing each characteristic on the treatment indicator and strata fixed effects. Next, we stored the most extreme of these t-statistics, and selected the randomization where this value is smallest.[17]

II.D.2. Randomization of community contests among treatment units

In addition to the randomization strategy described in II.D.1., after selecting the randomizationwith the greatest balance, we randomized all of the 49 treatment pairs into two arms: one group of GPs with community contests (24 GPs), and one group without those contests (25 GPs). Both treatment arms received the kits and related training. All pairs of control GPs remained untouched. This randomization for the GP contests was done in July 2019.

[14] Specifically, for each GP we calculated the average performance score among all students from the baseline's one-on-one test (see below). For logistical reasons, we could not use the paper-based tests to stratify GPs.

[15] We follow Athey and Imbens (2017), who suggest that a fully pairwise randomized trial (with a single treated and a single control school, per pair) may complicate use of regression-based methods to analyze randomized trials.

[16] There was one left-over ("misfit") GP in each district (as 49 is not divisible by four). We paired these two GPs, randomly assigning one to the intervention group and the other to the control group (cf. Carril 2017).

[17] See Bruhn and McKenzie (2009), who call this approach the "minmax method." We are well aware that high numbers of re-randomization can lead to analytic problems, especially if the re-randomization strategy is unknown. We follow Banerjee et al. (2020) by pre-specifying our strategy and choosing a conservative number (ten) of re-randomizations.

Figure I depicts the study schools by treatment status. In Section V.A.1. below, we use the study's baseline data to investigate whether the randomization strategy led to comparable groups.

III. Hypotheses, Outcomes, and Data

III.A. Primary hypothesis and main outcome of interest

Our main research hypothesis is that the program improves students' mathematics learning. We measure this outcome in two ways: (1) Student math scores on standardized tests; and (2) Student math scores on a one-on-one test of basic mathematical skills.

III.A.1. Standardized math tests

We administered three rounds of standardized math tests to the students, in all sampled schools, to obtain baseline, midline, and endline assessments. These paper-based tests were administered to students in groups.[18] Assessments have 30-35 multiple-choice type items and students had a one-hour time limit.[19]

Test items are mapped to the official state curriculum, but also include items one or two years below grade level. These items had been administered in similar contexts in India, for large-scale assessments. The assessments do not use questions from Akshara's internal item bank. From these previous administrations, we used item response theory (IRT)-based item characteristics to maximize the assessments' test information.[20]

[18] At baseline, we were concerned that weak students could not answer a paper-based test. Therefore, we administered a subset of seven items both orally (one-on-one) and as written items. We found no floor effects, and so our concerns were unwarranted. Results are available upon request. In subsequent rounds, we used only written (group) standardized tests.

[19] Students typically took about 45 minutes to complete each test.

[20] See Jacob and Rothstein (2016) for an accessible introduction to item response theory.

III.A.2. One-on-one tests of basic mathematical skills

Due to its salience among policy makers, we also administered the well-known "ASER" test of basic arithmetic skill (cf. ASER 2017)[21] to the full sample of students, at the same time as the written assessments. These tablet-based tests were administered by trained enumerators. One-on-one test administration took, at most, ten minutes per student. We followed ASER's standard grading procedures, which classify test takers into five progressive ability levels: beginner, recognition of single-digit numbers, recognition of two-digit numbers, two-digit subtraction (with borrowing), and three-digit by one-digit division.

III.A.3. Estimates of student ability

We estimate each student's ability using a two-parameter logistic (2PL) IRT model (Birnbaum 1968; Samejima 1973).[22] We used anchor items across test rounds (baseline, midline, endline) to allow for linking of estimates onto a common, continuous ability scale (Stocking and Lord 1983; Kolen and Brennan 2004). More specifically, we treat each ASER level as an additional mathematics item, but constrain the written item parameters to match those from a model that uses the written test items only.[23]

We describe in more detail the test design and related validity evidence in Appendix B. The analyses in Appendix B confirm that the tests did not display floor or ceiling effects. They also suggest that our test items discriminate well; low ability students have a much lower probability of correctly answering difficult items than high ability students. The tests also exhibit low levels of noise, in

[21] The ASER is a comprehensive household survey of rural India. For children between 3-16 years, it records enrolment status and tests basic reading and arithmetic skills using a common set of testing tools.

[22] A 3PL model did not converge. Following our registered report, we used a 2PL model.

[23] Our registered report did not discuss how to combine oral and written items. Constraining the written item parameters follows our pre-registered plan to calculate an IRT-based test score based on written items, but also incorporates the information from the oral test.

terms of both overall reliability (as per Cronbach's α) and their precision for test takers with a wide range of ability levels (as per the test information function).

III.B. Secondary hypotheses and related outcomes

We pre-specified three sets of secondary hypotheses, along with the program's Theory of Change. We describe their respective outcomes here. First, we investigate the program's impact on student learning along more fine-grained sub-competencies of mathematical skill. Second, we examine three areas of intermediate outcomes: whether the program improved instructional behaviors, whether it changed students' attitudes towards mathematics, and whether it increased community engagement and parental involvement. Third, we assess the program's implementation fidelity and its immediate outputs.

III.B.1. Measures of sub-competencies

The study's standardized tests group items along two sets of (more fine-grained) domains: content domains and cognitive domains.

The tests capture students' ability on four content domains: Number sense; whole number operations; shapes and geometry; and data display, measurement, and statistics. Each test item is mapped to one of these content domains. The tests also capture students' ability on two cognitive domains: Knowing; and reasoning and applying. Each test item is mapped to one of these two cognitive domains. To construct summary outcome measures for each of the six domains, we calculate the percentage of related test items a student answered correctly.

III.B.2. Intermediate outcomes

Measures of instructional behaviors. We used unannounced classroom observation visits to measure instructional quality, time-on-task, and instructional behaviors in treatment and control schools, after the implementation of the program. These classroom visits were scheduled to follow the study's sample of students—not a given mathematics teacher. Thus, we focused strictly on the instruction these students actually received, regardless of whether their teachers changed over time. We conducted one round of classroom observations in the first school year (June 2018 to May 2019), and three additional rounds in the second school year (June 2019 to May 2020).

More specifically, we used a novel, standardized classroom observation instrument, developed by the World Bank, called "*Teach*". We selected this instrument for its relevance to the program's Theory of Change, for the academic rigor used to assess its psychometric properties, and since it was constructed in (not merely transferred to) developing countries. We adhered to the instrument as closely as possible; however, we also piloted in, and contextualized it for, government schools in Karnataka. *Teach* focuses on three broad domains of instructional quality: Classroom Culture, Instruction, and Socio-emotional Skills; each domain is clearly mapped to respective behavioral markers. To construct summary outcome measures for domains and sub-domains, we follow *Teach*'s standard procedures, as documented by Molina et al. (2020).

Teach is our main measure of instructional behavior, yet we complement it with two ancillary data sources: Teacher surveys (during school visits) and surveys of sub-samples of students. We focus on teachers' self-reported awareness of activity-based teaching methods, and their use of collaborative pedagogy.[24]

[24] During student surveys, we also asked three questions designed to measure student-reported quality of instruction: (a) whether students have difficulty understanding explanations; (b) whether the teacher provides interesting tasks during class; and (c) whether, if there are doubts, the teacher explains concepts again. We also asked students three questions on their collaboration with peers: (a) whether students ask classmates for help; (b) the extent of student collaboration during math class; and (c) their level of collaboration on homework. As preregistered, and following Bacher-Hicks et al. 2019, we prefer classroom observation measures to student reports. Results are available upon request.

Measures of parental involvement and community engagement. The student interviews included a battery of questions on parental involvement in their child's math education. The teacher interviews elicited teachers' perceptions on parental involvement, including when they last communicated with a parent. We also sought, in interviews with the sub-sample of parents, to measure parents' involvement in their child's math education.

To gather information on community engagement, we asked all headmasters about activities of their schools' School Development and Monitoring Committee (SDMC), as well as about parents' meetings with teachers.[25] During our process monitoring rounds done in the second school year, we supplemented these data by interviewing a school's GP leader and block education officer (BEO).[26]

Measures of student attitudes towards mathematics. We used surveys of the sub-sample of interviewed students to measure children's attitudes towards mathematics learning. We administered a battery of four questions.[27] We generate a summary index from these four items by calculating their inverse-covariance-matrix-weighted average (following Anderson 2008).

III.B.3. Implementation fidelity and program outputs

We use primary and secondary data to track implementation fidelity in treatment schools.[28] We organize these data by the program's two main components:

[25] As per India's Right to Free and Compulsory Education Act 2009 (RTE) and the Karnataka Gram Panchayat School Development Monitoring Committees Model Sub-Ordinance 2006, SDMCs formalize community involvement in school management and school improvement efforts. See Vaijayanti and Mondal (2015) for more information on SDMCs in Karnataka.

[26] BEOs oversee the provision of primary and secondary education in a block. BEOs are responsible for many tasks, including human resource management, school inspections and monitoring, academic support, and community engagement. For additional information on BEOs, see Aiyar and Bhattacharya (2016).

[27] We asked whether the student: (a) enjoys learning math; (b) is made nervous by math; (c) finds math hard to understand; and (d) finds math harder than other subjects.

[28] See Sabarwal, Evans, and Marshal (2014) on the importance of measuring program take-up thoroughly.

Teacher training and additional teaching inputs, and community contests ("GP contests"). The following subsections describe these measures in greater detail.

Teaching inputs for activity-based instruction, and related training. The Akshara Foundation provided us with administrative records on teachers' participation in GKA training sessions. We augmented these data by asking all teachers about their participation in, and perception of, these trainings.

During school visits, we recorded teachers' self-reports on the availability, and their use of, GKA materials. In the teacher survey, we asked teachers whether they were trained on how to use the teaching and learning materials, the availability and usage of those materials, and their perceptions of the program. Information on the availability and use of the GKA teaching and learning materials was also obtained from classroom observations and the school survey.

Finally, we gathered administrative information on the Akshara Foundation's monitoring efforts and (on-site) teacher re-trainings. Akshara requires its field staff to document all school visits through a mobile app. We used this information to count, for each school, the number of school visits.

Community contests. The research team attended all community contests ("GP contests"). During the contests, we recorded student attendance, which we mapped to the study's sample of students (using unique student IDs). At each contest, the research team also recorded parents' attendance. In student surveys during school visits, we also asked the students whether they had participated in the GP contests.

III.C. Disentangling the effect of program components

It is important to disentangle the effects of the GP contests from the GKA program's other effects. For example, a recent learning-by-play intervention in Ghana had positive impacts only if parents were *not* involved (Wolf et al. 2019). Our secondary work therefore investigates treatment effects separately by whether the program includes the GP contest component.

III.D. Cost data

To estimate the cost-effectiveness of the GKA program's approach to increase students' math skills, we collected data on implementation costs from the Akshara Foundation and planned implementation costs from the Government of Karnataka (actual implementation costs were unavailable). Costs include those borne by the Akshara Foundation,[29] by the Government of Karnataka,[30] and community contributions to GP contests.[31] We also calculated the opportunity cost of time for parents who attended the GP contests, using the average hourly duration of a contest and the average hourly wages in the study area.

To convert all cost and impact calculations to their present value in USD, we set 2018 as our base year and used nominal exchange rates and annual GDP deflator inflation rates for the years 2018 to 2020.[32] We assume a discount rate of 12 percent, as suggested in Dhaliwal et al. (2013), using the social opportunity cost of capital (SOC) approach.

III.E. Additional covariates

We collected demographic information from students (including gender, birthdate and parents' name) to use as additional covariates and to facilitate tracking of them over the study's multiple rounds of data collection.

[29] These are *actual* expenses on: (i) training, monitoring and reporting; (ii) teaching and learning materials; (iii) program support; and (iv) general administration.

[30] These are *planned* expenses on: (i) GKA kits; (ii) training and monitoring; (iii) evaluation; and (iv) GKA kits used for training and for the resource repositories of academic support institutions, such as District Institutes of Education and Training (DIETs).

[31] These are the actual expenses to conduct the event, such as transport and food.

[32] To calculate annual inflation rates we use the U.S. implicit GDP deflators for December 2017, December 2018, December 2019, and April 2020.

We also acquired additional administrative information for each school at baseline. In particular, we obtained data from official school report cards from the District Information System for Education ("DISE"), as well as data on each school's village from India's 2011 Census.[33]

III.F. Data collection and timeline

Appendix Figure A3 depicts the study's timeline, including both program implementation and data collection. The data collection began in November 2018 with the baseline survey, followed by four rounds of process monitoring (February 2019, August 2019, November 2019, and December 2019), a midline assessment (September 2019), and an endline assessment (February 2020).[34]

IV. Empirical Strategy

IV.A. Statistical model

IV.A.1. Average effects

We use the following specification to estimate the GKA program's impacts:

$$Y_{isgr}^{t} = \alpha_r + \beta^t T_{gr} + \gamma^t Y_{isgr}^{t=0} + \boldsymbol{\delta' X}_{isgr}^{t=0} + \epsilon_{isgr}^{t} \tag{1}$$

where Y_{isgr}^{t} is the outcome of interest for student i in school s, GP g, and randomization stratum r, at time t. In our primary analysis, Y_{isgr}^{t} refers to test scores. In our secondary analyses, Y_{isgr}^{t} consists of: (a) measures of sub-competencies; and (b) potentially mediating variables. The α_r terms are randomization strata fixed effects, T_{gr} is the treatment dummy and ϵ_{isgr}^{t} is the

[33] We used GIS software to match each school's location to its respective village.

[34] Our data collection followed J-PAL South Asia's strict data collection procedures, including double-entry of paper-based tests, high-frequency checks of electronic forms, spot-checks, and weekly monitoring and debriefs for field staff (see Glennerster 2017; J-PAL 2017).

residual term. To increase precision, all specifications include $Y_{isgr}^{t=0}$ and $\boldsymbol{X}_{isgr}^{t=0}$ as covariates. Measured at baseline ($t = 0$), $Y_{isgr}^{t=0}$ is a student's initial outcome of interest; $\boldsymbol{X}_{isgr}^{t=0}$ is a vector of baseline controls selected by a LASSO procedure on student age, gender, school-level DISE data, and village-level census data (see Dhar, Jain, and Jayachandran 2020). The coefficient of interest, β^t, is the program's intent-to-treat (ITT) effect, for each follow-up round t.

IV.A.2. Effects by program component

In our secondary analyses we estimate the additional effect of community contests by the following specification:

$$Y_{isgr}^{t} = \alpha_r + \beta_1^t T_{gr} + \beta_2^t D_{gr} + \gamma^t Y_{isgr}^{t=0} + \boldsymbol{\delta'} \boldsymbol{X}_{isgr}^{t=0} + \epsilon_{isgr}^{t} \tag{2}$$

where D_{gr} is a dummy indicating the treatment GPs randomly assigned to community contests, and all else is as in Equation (1). Thus, β_2^t indicates whether treatment effects are equal without or with the contests. We test for whether β_1^t alone, or the sum of β_1^t and β_2^t, is zero (the ITT effects of the program without and with the GP contests, respectively).

IV.A.3. Heterogeneous effects

We also use specifications that allow for heterogeneous treatment effects, by interacting potential moderators with the treatment indicator. We illustrate this with the specification for sub-group analysis by gender:

$$Y_{isgr}^{t} = \alpha_r + \beta^t T_{gr} + \theta^t T_{gr} * F_{isgr}^{t=0} + \zeta^t F_{isgr}^{t=0} + \gamma^t Y_{isgr}^{t=0} + \boldsymbol{\delta'} \boldsymbol{X}_{isgr}^{t=0} + \epsilon_{isgr}^{t} \tag{3}$$

Here, $F_{isgr}^{t=0}$ is the moderating variable of interest (in this case, a student gender indicator), measured at baseline, and all else is as defined above.

To avoid specification searching, we limit our heterogeneity analysis to three prespecified moderators: 1. Gender; 2. Initial level of ability; and 3. District.

IV.B. Statistical methods

IV.B.1. Estimation

We estimate standard OLS regressions; for the ASER data, which we use to create binary outcomes, we estimate linear probability models. We cluster standard errors at the GP level (cf. Abadie et al. 2017).

To check robustness, we use randomization inference to assess whether our re-randomization procedure led to unexpected consequences (Young 2019). In particular, we replicate our procedure for each of 5,000 iterations (cf. Heß 2017).

IV.B.2. Non-compliance

Lack of take-up. Schools and teachers may not take up the GKA program. We posit that the policy-relevant question is whether the program led to learning gains even for a (potentially) diluted treatment exposure. Our study thus estimates intent-to-treat (ITT) effects. Yet, we also report on the effectively observed program exposure,[35] and report on program outputs (see Section III.B.).

Spill-overs. We randomized at the GP level; we thus include multiple schools per randomization unit. Therefore, we expect no spill-overs from treatment to control schools. Yet, our school visits tracked schools' potential exposure to other, similar interventions (in both groups of schools). In particular, the ("*Nali Kali*") program, which promotes activity-based instruction in the lower grades, has been implemented in Karnataka. Yet, there is no overlap between

[35] In the experimental literature, som authors use "exposure" and "dosage" interchangeably. We prefer the term "observed exposure" to clearly distinguish subjects' effectively experienced treatment levels from their initially intended treatment levels.

this program and the grade levels investigated in our research.

IV.B.3. Missing values and attrition

We pre-registered strategies to address two types of missing values. Observations may contain incomplete data ("missing data"), or may not be observed in a later data-collection round ("attrition").

Missing data for observed observations. Students may leave individual test items blank. We decided to classify unanswered items as incorrect answers.

As with any nonequivalent anchor test (NEAT) design, students did not answer items that were not administered to them (i.e., questions not used as anchors; "missing by design"). In addition, a small share of students (3.7 percent) participated in only one of the two baseline tests (oral or written). The study's IRT models account for these missing values by using concurrent calibration, via marginal maximum likelihood estimation (Kolen and Brennan 2004).[36]

Attrition. We investigate attrition in two ways. First, we check whether it is systematically related to treatment status, through tests of differential attrition rates and of selective attrition.[37] Second, we employ two robustness checks: inverse-probability weighting (IPW) and Lee (2009) bounds.[38]

[36] We dropped students who took only the oral test and not the written baseline test. We retained those who took only the written test.

[37] Attrition is differential if it systematically differs across the treatment and control groups. It is selective when the mean of baseline test scores differs, conditional on treatment status (see Ghanem, Hirshleifer, and Ortiz-Becerra 2020).

[38] Third, if entire schools had attrited, we would have investigated robustness to dropping every school in those schools' randomization strata. Fortunately, each assessment round includes students from all 292 schools. Yet, we also present robustness checks for the subset of complete strata, dropping all schools in the strata of the two schools with zero enrollment.

IV.B.4. Multiple outcome and multiple hypothesis testing

We account for multiple hypothesis testing by using a summary measure of student learning as the primary outcome of interest. We interpret it as a "family" measure of math ability, akin to methods that use summary indices to adjust for multiple hypothesis testing (Anderson 2008; Kling, Liebman, and Katz 2007). Thus, we do not apply corrections to p-values (as in Romano and Wolf (2005) or Westfall and Young (1993); see List, Shaikh, and Xu (2019)).

V. Results

V.A. Internal validity, compliance, and program take-up

V.A.1. Attrition and balance

As shown in Table I and Appendix Table A1, randomization led to three groups of schools that are balanced in terms of observable student characteristics at baseline. Of the 78 comparisons across the three experimental groups, we detect only four statistically significant differences at the 5-percent significance level, which is well in line with what can be expected by chance. The main outcome variable (students' overall math score) is also balanced at baseline across the three groups.

The overall attrition rate from baseline to midline is 30 percent for the control group, and it is 21 percent from baseline to endline. Attrition from baseline to midline is not systematically different across experimental groups. At endline, attrition is slightly higher in the experimental group with community contests (by 3.2 percentage points), in comparison to the control group. However, as shown in Appendix Table A1, the non-attriting sample continues to be balanced on observable characteristics, across all three groups, both at midline and at endline.

V.A.2. Compliance and program take-up

We observed virtually full compliance of GPs' and schools' random assignment to treatment arms, with the exception of just one non-contest GP that received a contest. As shown in Figure A3, the one-week teacher training took place in January 2019, with a one-week refresher training provided in June 2019. Between the initial training and the midline survey, we estimate an exposure of 19 weeks. The exposure until endline was 37 weeks. Our calculations indicate that the effective number of working days over the study period was 215 days.[39]

Figure II summarizes the main indicators of implementation fidelity and program take-up over the study period. We consider three dimensions of analysis: (i) training, and teacher perception of the program; (ii) teaching inputs and take-up of materials; and (iii) community contests. In summary, although there are dimensions that can be improved in the future, we find that the program was largely implemented as intended.

For the training and perception dimension, we use both a headmaster and a teacher survey. The headmaster survey shows a high take-up rate: 96 percent of the treated schools actually participated in the GKA program, whereas none of the control schools did. Participation in any training and workshops since 2017 was high for both treated (99 percent) and control (93 percent) schools, according to our fourth and last teacher survey.[40] However, specific GKA training was received by 86 percent of fourth-grade math teachers, with no GKA trainings administered to control-group teachers. Similarly, 89 percent of teachers in treated schools reported having received training on how to use the GKA kits. As for on-site follow-up training and monitoring, NGO staff reported visiting 98 percent of the treatment schools at least once, and 82 percent of the treatment

[39] This number is based on the official school calendar, but removes any days with school closures (e.g., due to local festivals and holidays, or due to floods).

[40] Recall that GKA trainings replace the existing government training schedule; therefore, we do not expect a large difference in the percentage of teachers receiving *any* type of training.

schools at least twice over the study period (they did not visit control schools). Overall, 81 percent of math teachers in treated schools perceived that the GKA program had a large impact.

We report on seven indicators related to teaching inputs and take-up of materials. Almost all (94 percent) teachers reported having received the GKA kit. Teachers in treated schools also reported conducting group activities more frequently: 56 percent of them do so two to three times per week, compared to 40 percent in control schools. The difference is even larger for group activities conducted every day, at 38 percent for treated schools and 15 percent for control schools. About 60 percent of treatment-group teachers reported using the GKA kit for math classes every class. Classroom observations using the *Teach* instrument reveal that 41 percent of teachers in treated schools conduct group activities during class. This is a 30 percentage-point difference with respect to control schools. While 13 percent of teachers in control schools used teaching and learning materials (TLMs) in class, the proportion is considerably larger for teachers in treated schools (75 percent). In almost all of these cases when a treatment teacher used teaching and learning materials, the TLMs had been provided by the GKA program (72 percent overall, or 96 percent of the treatment-group teachers who used TLMs).

Finally, we investigate whether community contests were implemented as intended. The GP contests took place between August 2019 and January 2020, and there were 24 days on which contests were held. Here, we focus on schools assigned to the kit-plus-contests treatment arm (in comparison to control-group schools). The GP contest survey shows that 86 percent of the 71 treated-with-contests schools participated in the GP contests. The headmaster survey indicates that 33 percent of the schools participating in the contests received a report card after the contest. Our parent survey suggests that 11 percent of parents attended the GP contests. Our last indicator uses GP contest data, which shows that 73 percent of students participated in the contests.

V.B. Main results

Panel A of Table II summarizes the study's main results. In the time period from baseline to midline, control-school students' math scores improved by 0.13 standard deviations (statistically insignificant). In the time period from baseline to endline, control-school students' math scores improved by 0.40 standard deviations ($p < 0.01$). At both midline and endline, the difference across students in treatment schools and control schools is statistically indistinguishable from zero ($p > 0.1$)—that is, conditional on the vector of covariates, we cannot reject that treatment school students learned an equal amount when compared to their peers in control group schools.

V.C. Secondary results

V.C.1. Results for additional assessment outcomes

The remaining panels of Table II provide secondary results. First, we investigate the proportion of students who mastered each learning level of the ASER arithmetic test. In the time period from baseline to midline, the proportion of control school students who recognized two-digit numbers increased by 6 percentage points, from a base level of 90 percent. The percentage of students at the "subtraction" level increased by 9 percentage points, from a base level of 35 percent. At endline, improvements are more pronounced and they also include a 11 percentage-point increase in the proportion of students who know division (from a base level of 10 percent). However, at midline and at endline the difference in learning levels across students in treatment schools and control schools is close to zero and statistically insignificant—that is, treatment-group students did not perform differently on the ASER test, in comparison to their peers in control-group schools.

Next, in Panel C, we compare continuous test-scores for those written test items that are mapped to higher-order thinking skills ("HOTS") vs those mapped to lower-order thinking skills ("LOTS"). We find neither positive nor negative effects after eight months of the intervention. At endline, 13 months after the start of the intervention, the impacts for HOTS and LOTS are statistically indistinguishable, but we find a marginally significant impact of 0.11 SDs for lower-order thinking skills only ($p < 0.1$).

Lastly, in Panel D, we investigate the percentage of written questions students solved correctly, for each of the four mathematical content domains (data, geometry, number sense, and whole number operations).[41] At midline, we do not detect notable differences across the treatment and control group students, for any of the four sub-domains. At endline, we document how treatment student's proportion of correctly answered geometry questions increased by 4 percentage points ($p < 0.01$), with no statistically significant impacts for the remaining three content domains.

V.C.2. Results by program variant

We repeat the above analyses to investigate different impacts across the two types of treatment groups: (1) the group of schools with the full intervention, including Gram Panchayat contests, and (2) the group of schools that received the intervention without Gram Panchayat contests. Table III provides our results.

Concerning the study's main outcome, 13 months after the launch of the intervention, we find marginally significant, positive effects of the intervention

[41] Recall that our tests differed across assessment rounds. Therefore, these percentages of correctly answered questions are not comparable across rounds, and the reported gains should be interpreted with great caution. In contrast, our IRT-based test scores are reported on a common scale. Therefore, gains can be readily interpreted.

type without community contests (0.12 SDs, $p < 0.1$). These impacts are driven by positive effects on lower-order thinking skills (0.14 SDs, $p < 0.05$), and on geometry questions (5 percentage points, $p < 0.01$). At the same time, they are not reflected in the results of the ASER test. We do not find supportive evidence for the effectiveness of the full intervention that includes community contests. For both program variants, we do not find clear evidence for effects after 8 months.

V.C.3. Heterogeneous effects

We now investigate whether the effects on student math learning differ for three different subgroups of students. We provide results by gender, by students performance on the written baseline test (by tercile), and by district (Bijapur vs Tumkur). Table IV provides the pooled intention-to-treat effects on the study's main outcome measure. Table V provides these results by intervention type.

Table IV and Table V show that positive program effects are entirely driven by improvements among girls. For girls, we find marginally significant improvements of 0.14 SDs overall ($p < 0.1$), and of 0.18 SDs for the intervention type without community contests ($p < 0.05$). In contrast, for boys, relative to girls, these effects are 0.16 SDs lower overall ($p < 0.05$) and 0.15 SDs lower for the intervention type without community contests ($p < 0.1$). That is, for boys, coefficients are very close to zero and they are statistically insignificant. We also report a 0.17 SD difference in effect sizes for the intervention type with contests, favoring girls ($p < 0.05$). We do not observe clear patterns of heterogeneous effects for the remaining two subgroups of students.

V.C.4. Results for intermediate outcomes

Measures of instructional behaviors. In Figure III, we present the program's effects on teaching quality, for each of the two program variants.[42] As shown in the top panel, for the version with community contests, we do not detect a statistically significant impact on the overall quality of teaching as measured by the *Teach* index ($p > 0.1$). This null finding for the overall index masks a positive impact of 0.24 SDs on the dimension of teaching that is expected to promote students' socioemotional skills (e.g., whether the teacher promotes collaborative skills) and a negative effect of 0.14 SDs on the dimension of teaching related to classroom culture (e.g., whether the teacher creates a supportive learning environment). We do not observe impacts on the instruction dimension (e.g., whether the teacher provides high-quality feedback).

For the version without community contests, however, we document a positive effect of 0.11 SDs on the overall index of teaching quality, but we cannot rule out with confidence that the coefficient is in fact different from zero ($p > 0.1$). This overall finding is driven by a 0.17 SD improvements for the subdimension related to socioemotional skills ($p < 0.05$). We document a (statistically insignificant) 0.08 SD improvement in instructional quality. Once community contests are removed from the program, the coefficient for the dimension of classroom culture is close to zero and statistically insignificant.

As shown in Figure III's bottom panel, the negative effects of adding community contests to the intervention coincide with the intervention timeline (compare to Appendix Figure A3). Prior to the contests, instruction across the treatment groups was indistinguishable (data collection Round 1). The launch of contests approximately coincided with data collection Round 2. Thereafter,

[42] Observers also recorded whether teachers spent their time "on task" and whether the observed math class appeared staged. We do not observe any differences across the treatment and control groups on these indicators.

in Rounds 3 and 4, we find large negative effects on the overall quality of instruction (between 0.26 and .30 SDs), and in particular on the dimension of classroom culture (between 0.30 and 0.48 SDs).

Measures of parental involvement and student attitudes. In Figure IV we present the program's impacts on parental involvement and student attitudes, for each of the two program types. From the top panel, we cannot conclude that the version with community contests increased parent-reported parental engagement (e.g., how often parents sit with their child to supervise their math homework). In addition, teacher-reported parental engagement did not improve (e.g., how often parents reach out to teachers to discuss their child's performance). Similarly, for the version without contests, both parent- and teacher-reported parental involvement does not differ in comparison to control schools.

In the bottom panel of Figure IV, we document overall positive effects on the study's index of student attitudes towards mathematics (e.g., whether students enjoy the subject or, in contrast, whether mathematics makes them nervous). The point estimates are similar (0.09 SD overall; $p < 0.05$), with a 0.10 SD improvement for the full intervention including contests, and a 0.08 SD improvement for the version without. However, the pooled estimate is more precise; the coefficient for the version without contests loses its statistical significance as impacts are estimated separately, by program variant.

V.D. Robustness checks

As discussed earlier, we subject the study's main findings to a series of robustness checks. Table VI presents their results. The outcome of interest is students' overall math score at endline, standardized with respect to the control group at baseline.

Columns (1) to (3) investigate robustness to attrition. We present inverse probability-weighted estimates and Lee (2009) bounds. Columns (4) to (7) investigate robustness to alternative sample definitions. We present results for the study sample of students conditional on participating in any baseline test, the study sample of students conditional on participating in any baseline test and at least a written endline test (but not necessarily an oral endline test), a sample where we remove a randomization stratum with contamination (one treatment school in the group without community contests received a contest), and a sample of schools that drop those strata where a school was dropped after baseline (two schools had zero attendance at baseline). Column (8) investigates robustness to measurement decisions. We present results for an outcome measure that uses written test items only, ignoring students' performance on the oral test. Finally, we investigate robustness to the re-randomization procedure used for treatment assignment. Column (9) presents randomization inference (RI)-based p-values, where we repeated the same randomization procedure in each of 5,000 RI iterations.

In general, our point estimates remain remarkably similar across all robustness checks; we are confident that they are not substantially affected by attrition, the study's sample definition, or our choices in constructing the summary measure of mathematics learning. However, the precision of our findings is somewhat reduced for the randomization-inference based estimate: The p-value on the ITT effect for the program without contests increases to 0.27; the one for the same effect among girls increases to 0.15. This reduction in the statistical significance of our results when randomization inference is used should be interpreted with caution because, as explained in Athey and Imbens (2017, p. 89), the sampling variance for the estimated average treatment effect that is calculated using randomization inference omits the sampling variance of unit-level treatment effects (since it is not possible to estimate the latter variance consistently). Since this latter variance reduces the overall variance of the estimated average treatment effect, omitting it overestimates the overall variance of the average treatment effect obtained by using randomization inference. Even so, we continue to see strong evidence for our conclusion that the addition of community contests did not lead to improvements in mathematics learning, over and beyond the program variant without the contests.

V.E. Program costs

We calculate program costs using the present value streams of the cost data described in Section III.D., over the time period from baseline to endline (approximately 15 months). Costs include those borne by the Akshara Foundation (USD 1,258 for all treated schools), by the Government of Karnataka (USD 10,000 for all treated schools),[43] and the community contribution to GP contests (USD 1,245 for all treated schools).[44] The calculated opportunity cost to all parents who attended the GP contests is USD 22.[45] The number of treatment-group students in the study sample is 2,059.

Across the two program versions, the average program cost is USD 7.4 per student. The variant of the program without GP contests had a cost of USD 6.8 per student; the variant with the GP contests was USD 8 per student. Our results precisely rule out additional benefits of adding community contests and, therefore, it is more cost effective to remove the contests.

In comparison, the program's costs are substantially lower than the per-

[43] The costs of the kits were calculated based on the assumption that it takes three years for them to fully depreciate.

[44] Students in grades 4, 5 and 6 participated in the GP contests, but we do not have actual costs per grade. Thus, to calculate the approximate costs incurred in having our treatment group participating in the contests, we used the 2017 enrollment rates from DISE. We calculated the proportion of students in 5th grade among those enrolled in grades 4, 5 and 6, and multiplied the total GP contests costs by this proportion.

[45] Only 14 parents from our treatment group attended the GP contests.

student costs required for a successful intervention that was recently evaluated in Indian government schools, in Rajasthan (Muralidharan and Singh 2019). More specifically, the Rajasthan program increased mathematics test scores by 0.21 standard deviations, at a cost of about USD 66 per student per year. In comparison, if all costs from our program are attributed to girls (given the null findings among boys), it increases their test scores by 0.18 standard deviations at a cost of USD 14.

VI. Conclusion

To achieve increased economic growth and, more generally, a higher quality of life, many developing countries have substantially increased spending on education. This has led to large increases in enrollment in primary (and secondary) education, yet these positive educational outcomes are unlikely to lead to higher economic growth and improvements in the quality of life if students learn much less than the curriculum expects them to master. The academic performance of primary school students in many developing countries is disappointingly low, and India is one of those countries. This state of affairs has opened a serious debate on "what works" to improve learning outcomes in developing countries. Recent reviews of the literature point to pedagogical interventions and teacher training as among the most effective education interventions to increase student learning. Yet, these findings draw upon a very small evidence base.

Our research adds to this small but important evidence base. We estimate the causal effects of an innovative program in the state of Karnataka, India, that promotes activity-based learning of mathematics at the primary school level through additional teaching inputs, related teacher training, and community engagement. The "Ganitha Kalika Andolana" (GKA) program is designed to help students learn mathematical concepts and to develop their concrete mathematical understanding through engaging activities (before moving on to representational and abstract learning)—in contrast with the conventional chalk-and-talk method commonly used in Indian schools.

To estimate the causal effect of this program on student learning in mathematics, we implemented a randomized controlled trial in 98 administrative units (Gram Panchayats), dividing these units, and 292 schools within them, into either the program group or a control group. To assess whether community contests added to the program's effectiveness, we moreover randomly assigned the treatment group into two sets of Gram Panchayats. One set received the full program version that included community contests, and the other received a version that excluded the contests.

Our analysis shows adherence to this study design, and that the program was largely implemented as intended. More specifically, 86 percent of grade-4 teachers in the program schools received the GKA training, all program schools received the additional teaching inputs (GKA kits), there were large differences in the pedagogical methods used by the program school teachers (relative to control group teachers), and, in the group assigned to the full intervention, 86 percent of the program schools participated in the community contests (GP contests). Analyses of balance across experimental groups in terms of their observable characteristics (before and after attrition) lend additional evidence for the study's internal validity.

Our primary outcome of interest is learning in mathematics among grade 4 students, as measured by both oral and written mathematics assessments. Thirteen months after the launch of the intervention, we find that, on average, the program had at most small (0.07 standard deviations), statistically insignificant impacts on fourth-grade students' learning of mathematics. Analysis by gender finds a marginally significant impact of 0.14 standard deviations for girls' math scores, but no effect for boys. Analysis by program type reveals marginally significant, positive effects for the version of the program that does not include community contests (0.12 standard deviations), with significant effects of 0.18 standard deviations among girls (and close-to-zero effects among boys) for that version of the program.

Differences in the impacts of the intervention with and without the community contests are surprising and potentially important for policy. We find marginally significant, positive effects of the intervention variant without community contests on grade 4 students' learning of mathematics (0.12 SDs, $p<0.1$), but essentially no effect of the intervention variant that includes these contests. Consistent with this, we find little effect on overall teaching quality of the program variant that includes the contests, while we find a positive effect of 0.11 SDs on the overall index of teaching quality for the variant without the contests, although we cannot rule out that the coefficient is in fact different from zero ($p>0.1$). This overall finding is driven by a 0.17 SD improvement on teaching practices that are expected to promote socioemotional skills ($p<0.05$).

We document additional evidence of disappointing effects of the program variant with the contests on the quality of instruction. Prior to or at the time of the contests, instruction quality across the two treatment groups was indistinguishable, but after the contests there are large negative effects on the overall quality of instruction in the schools with the contests (relative to schools without the contests). These negative effects are driven by students' exposure to a less hospitable classroom environment (-0.14 SDs, for the overall study period; -0.30 to -0.48 SDs, in the remaining study period after the contests had been conducted). The apparent negative impacts of the contests suggest that efforts to increase community engagement may "backfire" if they trigger undesirable behaviors in public servants.

Future research could go in several directions. First, it is possible that program impacts would be larger (or smaller) when implemented over a period longer than the 13-month intervention period covered in this evaluation. Second, research on skills other than mathematics, such as reading and science skills, would be highly informative. Third, the differences in program effects by gender merit further investigation, which would probably require a larger sample and more classroom observation data. Fourth, given the high proportion of primary-school students in India who are enrolled in private schools, an evaluation of this program's effectiveness in private schools would appear to be very valuable. Finally, the negative impact of the community contests on classroom culture implies that further research is needed on what types of pressure, when applied to public servants, may lead to unintended negative consequences.

Massachusetts Institute of Technology
United Nations Development Programme
University of Minnesota
University of Minnesota

Appendix

A Additional figures and tables

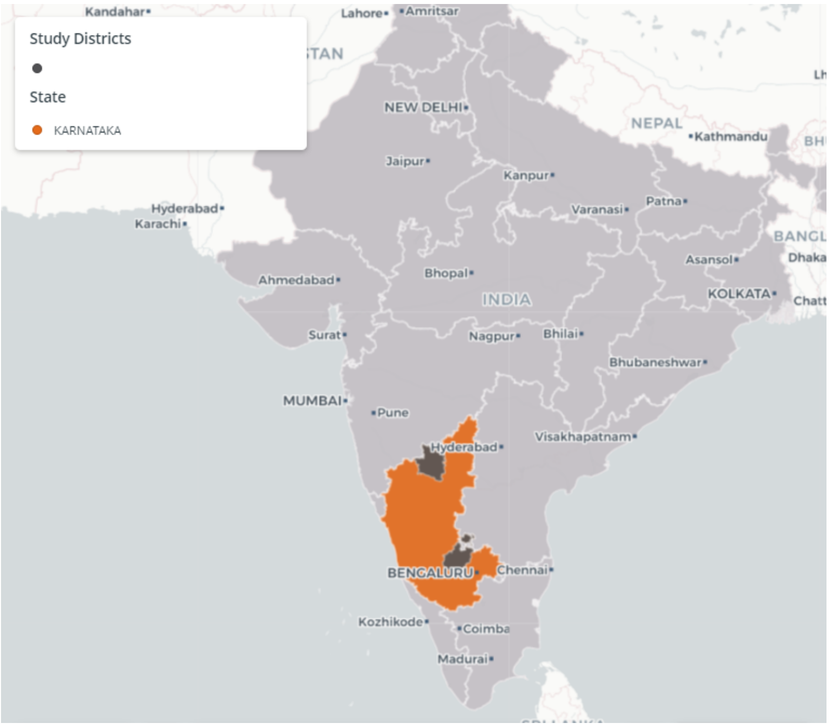


Figure A1
Location of the study

This figure depicts the state of Karnataka and the two districts selected for the study (Bijapur in the North and Tumkur in the South).

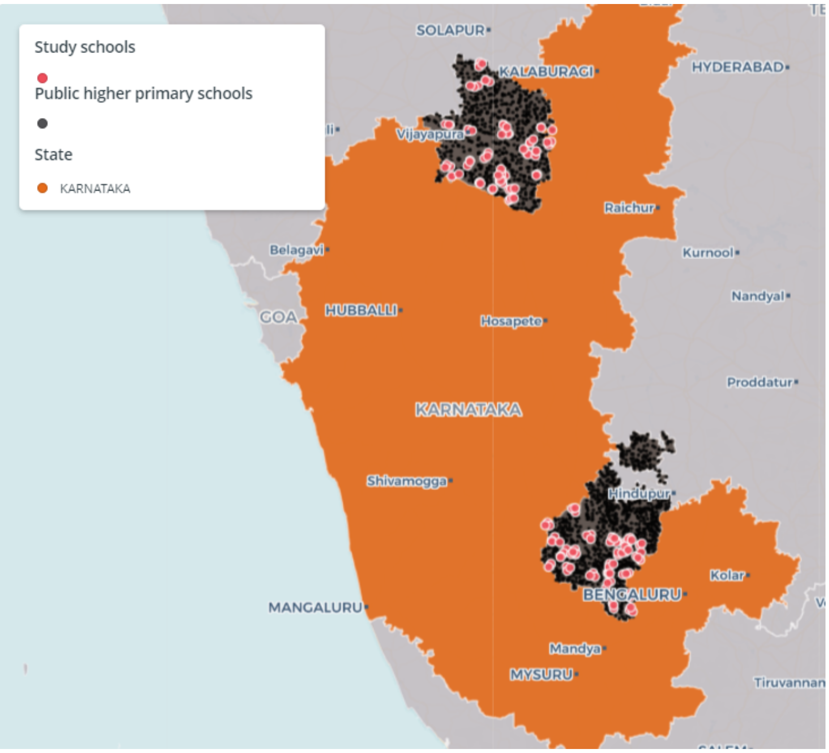


Figure A2
Random sampling of GPs and primary schools in study districts

This figure depicts all public higher primary schools in the study districts. Randomly selected schools in red. Sampling followed a two-stage procedure. First, we selected 98 GPs with at least 3 schools with grade-4 enrollment of 5 or more (49 per district; probability proportional to GPs' number of schools). Second, we selected 3 schools per GP at random.

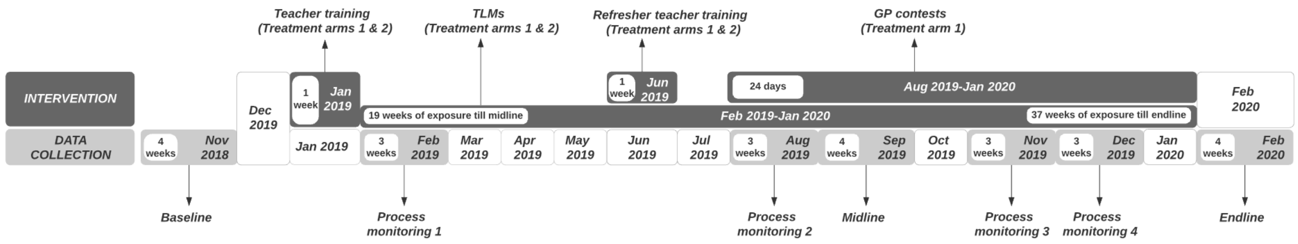


Figure A3
Study timeline

This figure depicts the study's timeline. Program implementation activities shown at the top, in dark gray. Data collection activities shown below, in light gray. "TLMs" stands for teaching and learning materials. Randomization of treatment-group GPs to the variant with vs without contests occurred in July 2019, after the first round of process monitoring had been completed.

TABLE A1
Non-attritor characteristics at baseline

	Number of observations			Mean			Differences		
	Control	Contests	Materials	Control	Contests	Materials	Contests vs Control	Materials vs Control	Contests vs Materials
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student age (as of 31-Dec-18)	1464	753	805	9.13	9.13	9.15	-0.01	0.02	-0.04
				[0.54]	[0.54]	[0.58]	(0.03)	(0.03)	(0.03)
Female	1472	755	812	0.55	0.55	0.54	0.00	-0.01	0.01
				[0.50]	[0.50]	[0.50]	(0.03)	(0.03)	(0.03)
Math Score (2pl, std.)	1472	755	812	0.05	-0.01	-0.02	-0.02	-0.08	0.06
				[0.97]	[0.97]	[0.96]	(0.05)	(0.08)	(0.07)
ASER (Baseline)>=1-digit	1472	755	812	0.99	0.99	0.99	-0.01	-0.01	-0.00
				[0.09]	[0.11]	[0.11]	(0.01)	(0.00)	(0.01)
ASER (Baseline)>=2-digit	1472	755	812	0.90	0.88	0.92	-0.02	0.02	-0.03**
				[0.30]	[0.32]	[0.27]	(0.02)	(0.01)	(0.02)
ASER (Baseline)>=Subtraction	1472	755	812	0.34	0.34	0.34	-0.00	-0.01	0.01
				[0.48]	[0.47]	[0.48]	(0.02)	(0.02)	(0.02)
ASER (Baseline)>=Division	1472	755	812	0.09	0.10	0.10	0.01	0.01	0.00
				[0.29]	[0.30]	[0.30]	(0.01)	(0.02)	(0.02)
Math, HOTS (2pl, std.)	1472	755	812	0.05	-0.02	-0.03	-0.02	-0.08	0.06
				[0.98]	[0.99]	[0.98]	(0.06)	(0.09)	(0.09)
Math, LOTS (2pl, std.)	1472	755	812	0.05	-0.01	-0.03	-0.02	-0.08	0.06
				[0.97]	[0.96]	[0.96]	(0.05)	(0.07)	(0.07)
Data Percent Correct (Baseline)	1472	755	812	0.38	0.35	0.36	-0.02	-0.02	0.00
				[0.21]	[0.21]	[0.21]	(0.01)	(0.02)	(0.02)
Geometry Percent Correct (Baseline)	1472	755	812	0.49	0.48	0.48	0.00	-0.02	0.02
				[0.29]	[0.28]	[0.29]	(0.02)	(0.03)	(0.03)
Number Sense Percent Correct (Baseline)	1472	755	812	0.61	0.59	0.59	-0.01	-0.01	0.00
				[0.27]	[0.28]	[0.27]	(0.01)	(0.02)	(0.02)
Whole Number Operations Percent Correct (Baseline)	1472	755	812	0.53	0.52	0.50	0.00	-0.02	0.03
				[0.28]	[0.28]	[0.28]	(0.02)	(0.02)	(0.02)

Notes. This table provides descriptive statistics for the study sample, by treatment status. “Contests” refers to the full treatment; “Materials” refers to the treatment without contests. Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). “Non-attritor” refers to a student who took the baseline and endline assessments. All estimations include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

B Test design and validity evidence

We measure student achievement in mathematics with tests that seek to capture what students know and can do in this subject area, with direct reference to their schools' official Kannada-medium curriculum. The assessments are summative and of low stakes, both for the test takers and for the study's schools. These tests were administered under the supervision of the research team at baseline, midline, and endline. In this appendix, we present validity evidence for the tests' contents and for the tests' internal coherence as observed at baseline—results for the midline and endline assessments produce similar results (available upon request).

Content validity

The tests were administered on paper, as multiple-choice tests, and contained 32 items. Questions on the tests are mapped to four content areas (data display, geometric shapes and measures, number sense, and whole number operations), with eight questions per content area. Within each content area, half of the questions tap into higher-order thinking skills; the remaining half are associated with lower-order thinking skills. Overall, about 50 percent of items are mapped to students' enrolled grade level. The remaining 50 percent are mapped to curricular content from lower grades.

We further improved the test's content validity through four strategies, as follows. First, prior to the tests, we discussed the test blueprint and content with the implementing organization.[46] Secondly, for each round of assessments, we reviewed the test questions with an external panel of subject matter experts.[47] Third, we mapped each test question to the official schoolbooks used

[46] To ensure that the test administration remained impartial and unbiased, we did not repeat this strategy for the midline and endline tests.

[47] The panel consisted of former teachers and curriculum experts. The panel did not include

in Karnataka. Fourth, we accompanied each round of test development with (out-of-sample) field pilots, to further assess the local relevance of questions and their use of Kannada language.

Internal coherence and reliability

We begin our analysis of test coherence and reliability by investigating floor and ceiling effects. If all (or no) students were able to solve test questions correctly, we would not be able to distinguish students of different achievement levels. Figure B1 presents the mean percentage of correct responses for the baseline test (for all test questions, and by cognitive and content domains). It shows that, on average, students solved approximately half (48.5 percent) of the test questions correctly. Figure B2 presents the distribution of percentage of correct responses for the baseline test (again, for all test questions, and by cognitive and content domains). It shows that the distribution of test scores is approximately bell shaped, with no substantial "bumps" at the extremes of the performance distribution. Taken together, we find no evidence that floor or ceiling effects may limit the test's general validity.

Next, we turn to the *range* of ability covered by test questions. Table B1 displays the *a* and *b* parameters for the 32 test questions, as per a two parameter logistic (2PL) item response theory (IRT) model.[48] The table's *b* parameters show how the test offers a well-distributed measure of achievement in mathematics, as items cover a wide range of difficulty. In addition, all but one of the items show high levels of discrimination.[49] From this analysis, we conclude that our test scores are informative over a wide range of student ability in this

staff of the implementing organization.

[48] A three parameter (3PL) model did not converge for the baseline data.

[49] We kept the item with low discrimination (Q1140) in the baseline assessment. However, we did not repeat the item in our midline or endline assessments (i.e., it does not serve as an "anchor item").

setting.

We continue by investigating whether these item characteristics translate into high levels of internal consistency. A measure of internal consistency shows how closely related a set of items is as a group. The Cronbach's alpha (Cα) is a widely used measure of reliability in psychometric testing. The Cα is a function of the number of items in a test, the covariance between pairs of items, and the variance of the total score. The theoretical value of Cα varies from 0 to 1, with a rule of thumb of 0.7 or higher suggesting that the test is reliable. In this study, the Cα is 0.91 for the 32 written items. We thus conclude that our instrument is highly reliable overall.

This overall reliability level may nevertheless not translate into high levels of precision for the full range of test takers (as low-ability and high-ability are usually measured with higher levels of noise). Lastly, we therefore consider an additional measure of precision: the test information function (TIF). The information function tells how precisely each ability level is being estimated by a given IRT model, along with the corresponding standard error of measurement, for a given level of ability level θ. Figure B3 presents the TIF curve for this study and corresponding standard errors. We find a low standard error of measurement for a wide range of ability—even students two standard deviations below (or above) the median are assessed with a standard error below 0.45 (corresponding to reliability levels above 0.8, even at these more extreme levels of student ability).

TABLE B1
Item characteristics

Item	a (Discrimination)	b (Difficulty)
Number sense (Q1)	1.805	-1.448
Number sense (Q6)	1.608	-1.185
Whole number operations (Q1106)	1.549	-1.007
Geometric shapes and measures (Q9)	1.769	-0.853
Data display (Q22)	1.397	-0.74
Whole number operations (Q1102)	1.47	-0.701
Number sense (Q2010)	2.502	-0.587
Data display (Q21)	0.936	-0.469
Geometric shapes and measures (Q2006)	1.273	-0.413
Data display (Q41186)	2.349	-0.302
Number sense (Q1138)	1.719	-0.249
Whole number operations (Q1110)	1.581	-0.194
Whole number operations (Q1105)	1.647	-0.138
Whole number operations (Q1118)	2.348	-0.064
Geometric shapes and measures (Q2011)	1.602	-0.03
Number sense (Q5)	1.174	-0.008
Geometric shapes and measures (Q1126)	1.602	0.012
Number sense (Q8)	1.745	0.058
Geometric shapes and measures (Q2007)	1.921	0.086
Geometric shapes and measures (Q1162)	2.146	0.257
Whole number operations (Q1104)	1.73	0.32
Number sense (Q40)	1.022	0.41
Number sense (Q41)	1.669	0.442
Data display (Q2004)	1.529	0.519
Whole number operations (Q38)	1.613	0.52
Data display (Q30)	1.32	0.856
Geometric shapes and measures (Q1127)	1.036	1.104
Geometric shapes and measures (Q2002)	0.855	1.344
Number sense (Q25)	0.76	1.792
Data display (Q2008)	0.846	1.883
Data display (Q2001)	0.576	5.745
Data display (Q1140)	0.022	106.964

Notes: This table reports on items' discrimination and difficulty parameters, for the written baseline test as per a 2PL IRT model. Item numbers (in parentheses) refer to study-internal question IDs. Items are sorted by difficulty; items cover a wide range of difficulties. With the exception of one item (Q1140), items discriminate well.

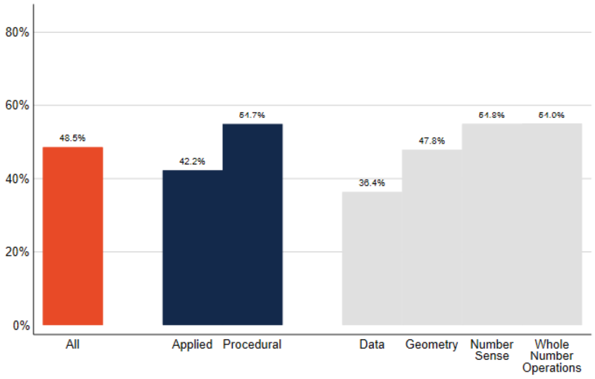


Figure B1
Mean percentage of items solved correctly (Baseline)

This figure provides the mean percentage of test questions students solved correctly during baseline (overall, by cognitive domains, and by content domains).

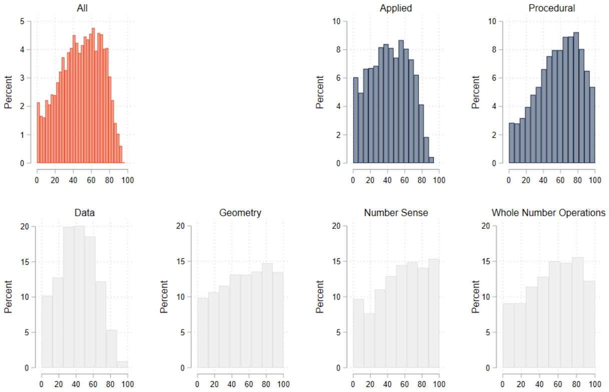


Figure B2
Distribution of percentage of items solved correctly (Baseline)

This figure provides histograms of the percentage of test questions students solved correctly during baseline (overall, by cognitive domains, and by content domains).

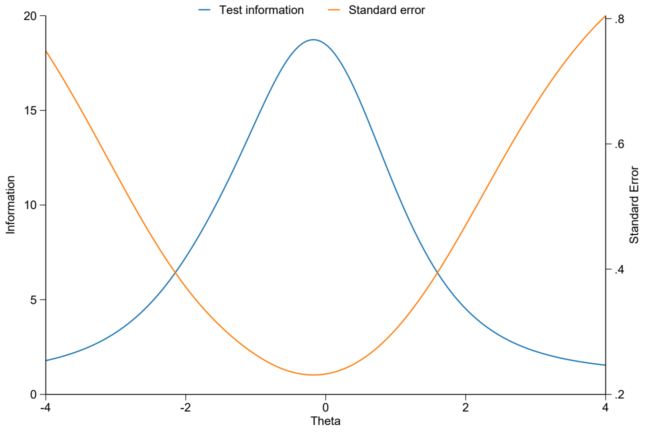


FIGURE B3

Test information function (TIF)

This figure provides the test information function, and corresponding standard errors of measurement, for the baseline as per a 2PL IRT model.

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* Technical report w24003. Cambridge, MA: National Bureau of Economic Research.

Acemoglu, Daron, Camilo García-Jimeno, and James A. Robinson. 2015. "State Capacity and Economic Development: A Network Approach." *American Economic Review* 105, no. 8 (August): 2364–2409.

Aiyar, Yamini, and Shrayana Bhattacharya. 2016. "The Post Office Paradox: A Case Study of the Block Level Education Bureaucracy." *Economic & Political Weekly* 51 (11): 61–69.

Aker, Jenny C., and Christopher Ksoll. 2019. "Call me educated: Evidence from a mobile phone experiment in Niger ." *Economics of Education Review* 72 (October): 239–257.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–1165.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–1495.

Armstrong, Robert. 2000. "Performance Management." In *Human Resource Management,* 69–84. Oxford, UK: Heinemann.

ASER. 2017. *Annual Status of Education Report 2016 (Rural).* Provisional Report. New Delhi: Pratham.

ASER. 2018. *Annual Status of Education Report 2017 (Rural).* Full Report. New Delhi: Pratham.

Athey, Susan, and Guido Imbens. 2017. “The Econometrics of Randomized Experiments.” In *Handbook of Economic Field Experiments,* edited by Abhijit Banerjee and Esther Duflo, 1:73–140. Elsevier.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. “An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys.” *Economics of Education Review* 73 (December): 101919.

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

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. “Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India.” *American Economic Journal: Economic Policy* 2, no. 1 (February): 1–30.

Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. “A Theory of Experimenters: Robustness, Randomization, and Balance.” *American Economic Review* 110, no. 4 (April): 1206–1230.

Barrera-Osorio, Felipe, Paul Gertler, Nozomi Nakajima, and Harry Patrinos. 2020. *Promoting Parental Involvement in Schools: Evidence From Two Randomized Experiments.* Technical report w28040. National Bureau of Economic Research, November.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2020. “The Glittering Prizes: Career Incentives and Bureaucrat Performance.” *Review of Economic Studies* 87, no. 2 (March): 626–655.

Best, Michael Carlos, Jonas Hjort, and David Szakonyi. 2017. *Individuals and Organizations as Sources of State Effectiveness.* Working Paper 23350. National Bureau of Economic Research, April.

Birnbaum, Allan. 1968. “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability.” In *Statistical Theories of Mental Test Scores,* 397–479. Reading, MA: Addison-Wesley.

Björkman, Martina, and Jakob Svensson. 2009. “Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda.” *Quarterly Journal of Economics* 124, no. 2 (May): 735–769.

Björkman Nyqvist, Martina, Damien de Walque, and Jakob Svensson. 2017. “Experimental Evidence on the Long-Run Impact of Community-Based Monitoring.” *American Economic Journal: Applied Economics* 9, no. 1 (January): 33–69.

Blimpo, Moussa P., David Evans, and Nathalie Lahire. 2015. *Parental human capital and effective school management : evidence from The Gambia.* Working Paper 7238. Washington, D.C.: The World Bank, April.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng’ang’a, and Justin Sandefur. 2018. “Experimental evidence on scaling up education reforms in Kenya.” *Journal of Public Economics* 168 (December): 1–20.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232.

Bruner, Jerome S., and Helen J. Kenney. 1965. "Representation and Mathematics Learning." *Monographs of the Society for Research in Child Development* 30 (1): 50–59.

Carril, Alvaro. 2017. "Dealing with Misfits in Random Treatment Assignment." *Stata Journal* 17 (3): 652–667.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotzé, Nompumelelo Mohohlwanex, Stephen Taylor, and Tshegofatso Thulare. 2020. *Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning in South Africa.* Working Paper 20/050. Oxford: RISE, November.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. 2020. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *Journal of Human Resources* 55 (3): 926–962.

Conn, Katharine M. 2017. "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations." *Review of Educational Research* 87 (5): 863–898.

DellaVigna, Stefano, and Elizabeth Linos. 2020. *RCTs to Scale: Comprehensive Evidence from Two Nudge Units.* Technical report. Working Paper, UC Berkeley, May.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2013. “Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries - A General Framework with Applications for Education.” In *Education Policy in Developing Countries,* edited by Paul Glewwe, 285–338. Chicago: The University of Chicago Press.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. 2020. *Reshaping Adolescents’ Gender Attitudes: Evidence from a School-Based Experiment in India.* Working Paper 25331. National Bureau of Economic Research.

Dixit, Avinash. 2002. “Incentives and Organizations in the Public Sector: An Interpretative Review.” *Journal of Human Resources* 37 (4): 696–727.

Duflo, Annie, Jessica Kiessel, and Adrienne Lucas. 2020. *External Validity: Four Models of Improving Student Achievement.* Technical report w27298. Cambridge, MA: National Bureau of Economic Research, June.

Duflo, Esther. 2020. “Field Experiments and the Practice of Policy.” *American Economic Review* 110, no. 7 (July): 1952–1973.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. “School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools.” *Journal of Public Economics* 123 (March): 92–110.

Ganimian, Alejandro J., and Richard J. Murnane. 2016. “Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations.” *Review of Educational Research* 86 (3): 719–755.

Gertler, Paul J., Harry Anthony Patrinos, and Marta Rubio-Codina. 2012. "Empowering parents to improve education: Evidence from rural Mexico." *Journal of Development Economics* 99, no. 1 (September): 68–79.

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Becerra. 2020. *Testing Attrition Bias in Field Experiments.* Working Paper 113. Berkeley: Center for Effective Global Action, University of California, March.

Glennerster, Rachel. 2017. "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency." In *Handbook of Economic Field Experiments,* edited by Abhijit Banerjee and Esther Duflo, 1:175–243. Elsevier.

Goodnight, Melissa Rae, and Savitri Bobde. 2018. "Missing children in educational research: investigating school-based versus household-based assessments in India." *Comparative Education* 54 (2): 225–249.

Heß, Simon. 2017. "Randomization Inference with Stata: A Guide and Software." *Stata Journal* 17, no. 3 (September): 630–651.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics & Organization* 7 (Special Issue): 24–52.

Islam, Asad. 2019. "Parent–teacher meetings and student outcomes: Evidence from a developing country." *European Economic Review* 111 (January): 273–304.

J-PAL. 2017. *J-PAL Research Protocols.*

Jacob, Brian, and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives* 30 (3): 85–108.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119.

Kolen, Michael J, and Robert L Brennan. 2004. *Test Equating, Scaling, and Linking.* 3rd. New York, NY: Springer.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–1102.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple hypothesis testing in experimental economics." *Experimental Economics* 22, no. 4 (December): 773–793.

Majerowicz, Stephanie, and Ricardo Montero. 2018. "Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru." Cambridge, MA.

MHA. 2012. "15th Census of India."

Migdal, Joel S. 1988. *Strong Societies and Weak States: State-society Relations and State Capabilities in the Third World.* Princeton University Press, November.

Ministry of Law and Justice. 2009. *Right of Children to Free and Compulsory Education Act, 2009,* August.

Molina, Ezequiel, Syeda Farwa Fatima, Andrew Dean Ho, Carolina Melo, Tracy Marie Wilichowski, and Adelle Pushparatnam. 2020. “Measuring the quality of teaching practices in primary schools: Assessing the validity of the Teach observation tool in Punjab, Pakistan.” *Teaching and Teacher Education* 96 (November): 103171.

Muralidharan, Karthik, and Paul Niehaus. 2017. “Experimentation at Scale.” *Journal of Economic Perspectives* 31 (4): 103–124.

Muralidharan, Karthik, and Abhijeet Singh. 2019. “Improving Schooling Productivity through Computer-Aided Personalization: Experimental Evidence from Rajasthan.” Washington, D.C.: RISE Annual Conference 2019, June.

NIEPA. 2018. *U-DISE Flash Statistics 2016-17.* New Delhi, India: National Institute of Educational Planning / Administration.

Pandey, Priyanka, Sangeeta Goyal, and Venkatesh Sundararaman. 2011. *Does Information Improve School Accountability? Results of a Large Randomized Trial.* Discussion Paper 49. Washington, D.C.: The World Bank, December.

Popova, Anna, Violeta Arancibia, and David K. Evans. 2016. *Training Teachers on the Job: What Works and How to Measure it.* Working Paper 7834. Washington, D.C.: The World Bank.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. “Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia.” *American Economic Journal: Applied Economics* 6, no. 2 (April): 105–126.

Pritchett, Lant, Michael Woolcock, and Matt Andrews. 2013. “Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation.” *Journal of Development Studies* 49, no. 1 (January): 1–18.

Raffler, Pia, Daniel N Posner, and Doug Parkerson. 2020. “Can Citizen Pressure Be Induced to Improve Public Service Provision?” Cambridge, MA, October.

Romano, Joseph P., and Michael Wolf. 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica* 73 (4): 1237–1282.

Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshal. 2014. *The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone.* Policy Research Working Paper 7021. The World Bank.

Samejima, Fumiko. 1973. “A Comment on Birnbaum’s Three-Parameter Logistic Model in the Latent Trait Theory.” *Psychometrika* 38 (2): 221–233.

Sexton, Renard. 2020. “Information, Participation and Elite Disillusionment: Evidence from a Field Experiment in Peru.” Atlanta, May.

Stocking, Martha L., and Frederic M. Lord. 1983. “Developing a Common Metric in Item Response Theory.” *Applied Psychological Measurement* 7 (2): 201–210.

UNESCO Institute for Statistics. 2018. *Data for the Sustainable Development Goals.*

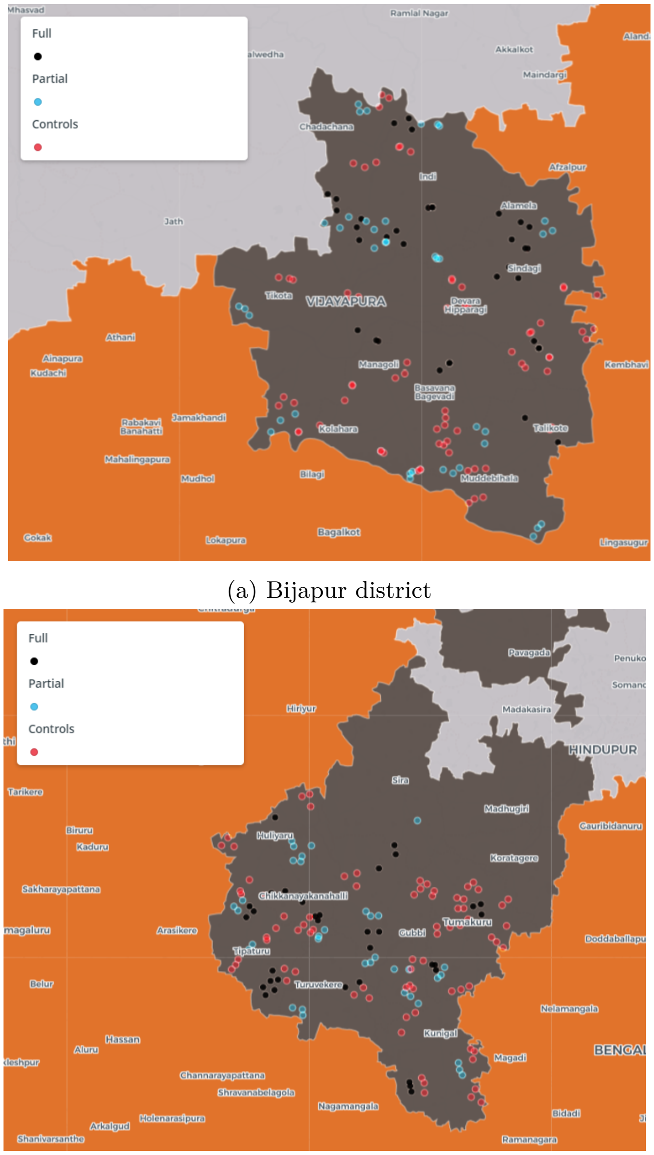
Vaijayanti, K., and Anuradha Mondal. 2015. “SDMCs in Karnataka: Analysing the quality of SDMC meetings in Hoskote, Kushtagi and Mundargi.” Bangalore.

Vivalt, Eva. 2020. “How Much Can We Generalize From Impact Evaluations?” *Journal of the European Economic Association* (September): 1–45.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-based multiple testing: examples and methods for P-value adjustment.* Wiley series in probability and mathematical statistics. New York: Wiley.

Wolf, Sharon, J. Lawrence Aber, Jere R. Behrman, and Edward Tsinigo. 2019. “Experimental Impacts of the “Quality Preschool for Ghana” Interventions on Teacher Professional Well-being, Classroom Quality, and Children’s School Readiness.” *Journal of Research on Educational Effectiveness* 12, no. 1 (January): 10–37.

Young, Alwyn. 2019. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” *Quarterly Journal of Economics* 134, no. 2 (May): 557–598.



(a) Bijapur district

(b) Tumkur district

Figure I

Study schools by treatment status

This figure depicts all study schools by treatment status. Stratified randomization at the GP-level within quadruplets of matched GPs. 50/50 treatment (T) and control (C), within districts, with subsequent randomization of treatment GPs, within strata, into T1 and T2. 10 re-randomizations to increase balance across T and C, following a "min-max" strategy (cf. Banerjee et al. 2020; Bruhn and McKenzie 2009).

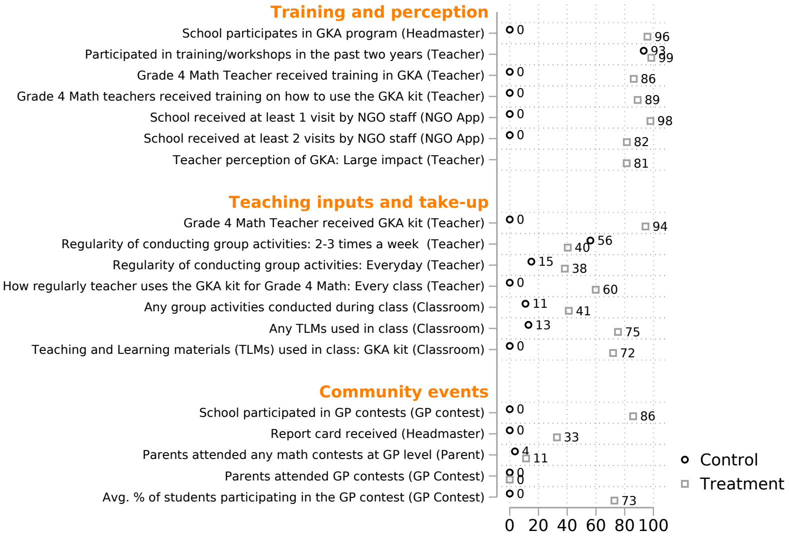
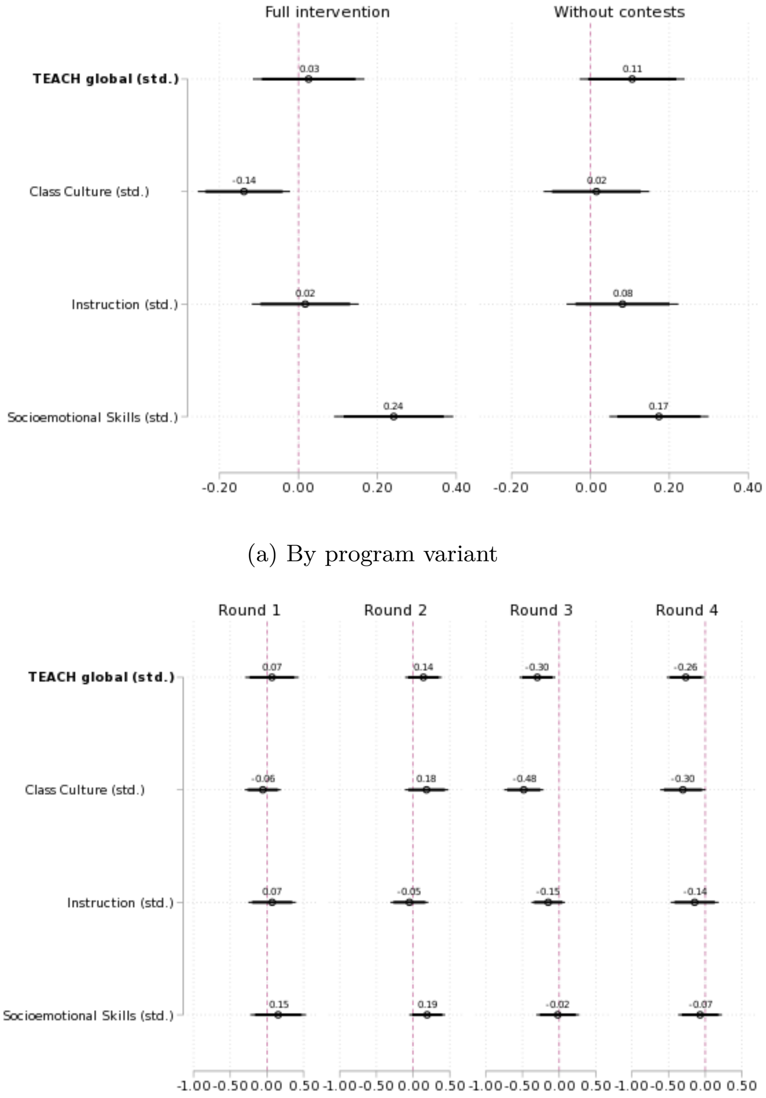


Figure II
Implementation fidelity and program take-up

This figure depicts the percentage of teachers or schools that satisfy indicators of implementation fidelity and program take-up, by experimental group. For the first two panels, "Treatment" includes schools in both treatment arms. The bottom panel excludes schools in the treatment arm without contests.

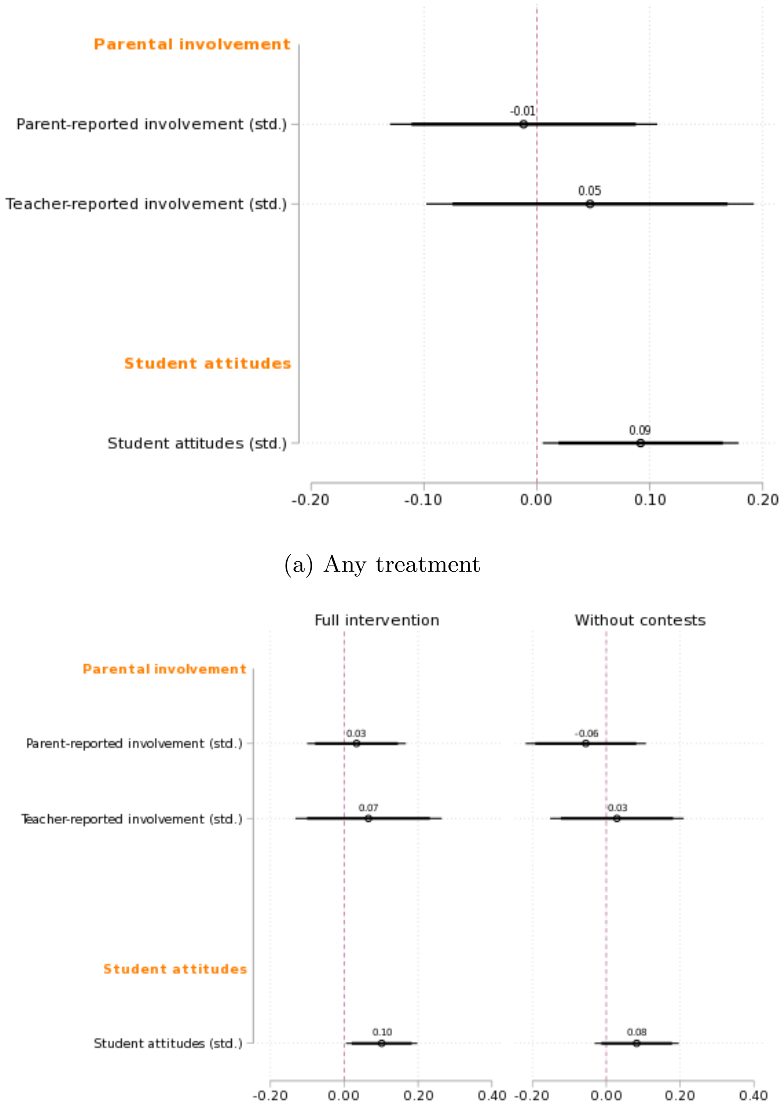


(a) By program variant

(b) Effects of adding community contests

Figure III
ITT effects on instructional quality

This figure depicts the program's ITT effects on instructional quality. Subfigure (a) shows ITT effects by program variant; subfigure (b) shows the effect of adding contests to the program (i.e., the difference across treatment groups), by data collection round. Randomization of treatment-group GPs to the variant with vs without contests occurred in July 2019, after Round 1 had been completed; the first contests started around the time of Round 2 (compare to the study timeline, depicted in Appendix Figure A3). The estimation sample consists of 1,615 classroom observation ratings. Thick/thin horizontal bars show 90-/95-percent confidence intervals.



(a) Any treatment

(b) By program variant

Figure IV
ITT effects on parental involvement, student attitudes

This figure depicts the program's ITT effects on parent- and teacher-reported parental involvement, as well as student attitudes towards math. Subfigure (a) shows overall results; Subfigure (b) shows results by program variant. The estimation samples consists of 1,937 parent interviews, 871 teacher interviews, and 5,649 student interviews, respectively. Thick/thin horizontal bars show 90-/95-percent confidence intervals.

TABLE I
Student characteristics at baseline

	Number of observations			Mean			Differences		
	Control	Contests	Materials	Control	Contests	Materials	Contests vs Control	Materials vs Control	Contests vs Materials
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student age (as of 31-Dec-18)	1852	999	1008	9.14	9.15	9.16	-0.00	0.03	-0.03
				[0.54]	[0.55]	[0.58]	(0.03)	(0.03)	(0.03)
Female	1862	1002	1017	0.53	0.53	0.52	0.00	-0.01	0.02
				[0.50]	[0.50]	[0.50]	(0.02)	(0.02)	(0.03)
Math Score (2pl, std.)	1862	1002	1017	0.01	-0.03	-0.07	-0.00	-0.07	0.07
				[0.99]	[0.96]	[0.98]	(0.05)	(0.07)	(0.07)
ASER (Baseline)>=1-digit	1862	1002	1017	0.99	0.98	0.98	-0.01	-0.01	-0.00
				[0.10]	[0.13]	[0.13]	(0.01)	(0.01)	(0.01)
ASER (Baseline)>=2-digit	1862	1002	1017	0.90	0.88	0.91	-0.02	0.01	-0.03**
				[0.30]	[0.33]	[0.28]	(0.02)	(0.01)	(0.01)
ASER (Baseline)>=Subtraction	1862	1002	1017	0.33	0.33	0.32	0.01	-0.02	0.02
				[0.47]	[0.47]	[0.47]	(0.02)	(0.02)	(0.02)
ASER (Baseline)>=Division	1862	1002	1017	0.09	0.09	0.10	0.00	0.01	-0.00
				[0.29]	[0.29]	[0.30]	(0.01)	(0.01)	(0.02)
Math, HOTS (2pl, std.)	1862	1002	1017	0.00	-0.03	-0.07	-0.00	-0.07	0.07
				[0.99]	[0.98]	[0.99]	(0.05)	(0.08)	(0.08)
Math, LOTS (2pl, std.)	1862	1002	1017	0.01	-0.02	-0.07	0.00	-0.08	0.08
				[0.99]	[0.95]	[0.98]	(0.05)	(0.07)	(0.06)
Data Percent Correct (Baseline)	1862	1002	1017	0.37	0.35	0.35	-0.02	-0.02	0.00
				[0.21]	[0.21]	[0.21]	(0.01)	(0.02)	(0.02)
Geometry Percent Correct (Baseline)	1862	1002	1017	0.48	0.48	0.46	0.00	-0.02	0.03
				[0.29]	[0.28]	[0.29]	(0.02)	(0.02)	(0.02)
Number Sense Percent Correct (Baseline)	1862	1002	1017	0.60	0.59	0.59	-0.00	-0.01	0.00
				[0.27]	[0.28]	[0.28]	(0.01)	(0.02)	(0.02)
Whole Number Operations Percent Correct (Baseline)	1862	1002	1017	0.52	0.52	0.49	0.01	-0.03	0.04**
				[0.29]	[0.28]	[0.28]	(0.01)	(0.02)	(0.02)
Attrition at midline	1862	1002	1017	0.30	0.30	0.30	-0.01	-0.01	0.00
				[0.46]	[0.46]	[0.46]	(0.02)	(0.02)	(0.03)
Attrition at endline	1862	1002	1017	0.21	0.25	0.20	0.03*	-0.01	0.04**
				[0.41]	[0.43]	[0.40]	(0.02)	(0.02)	(0.02)

Notes. This table provides descriptive statistics for the study sample, by treatment status. "Contests" refers to the full treatment; "Materials" refers to the treatment without contests. Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE II
ITT effects on student learning

	(1)	(2)	(3)	(4)	(5)
Panel A: Effects on main outcome					
Written test	0.08	0.13*	0.40***	-0.05	0.07
	[0.98]	(0.07)	(0.08)	(0.06)	(0.06)
Panel B: Effects on ASER test					
ASER>=1-digit	0.99	0.01	0.01***	-0.00**	-0.00
	[0.09]	(0.00)	(0.00)	(0.00)	(0.00)
ASER>=2-digit	0.90	0.06***	0.07***	-0.01	-0.01
	[0.29]	(0.01)	(0.01)	(0.01)	(0.01)
ASER>=Subtraction	0.35	0.09***	0.24***	-0.01	0.01
	[0.48]	(0.02)	(0.01)	(0.02)	(0.03)
ASER>=Division	0.10	0.01	0.11***	0.01	0.02
	[0.30]	(0.01)	(0.02)	(0.01)	(0.02)
Panel C: Effects by cognitive domain					
Higher-order	0.06	0.08	0.27***	-0.05	0.03
	[1.00]	(0.08)	(0.08)	(0.06)	(0.07)
Lower-order	0.09	0.10	0.31***	-0.05	0.11*
	[0.98]	(0.07)	(0.07)	(0.06)	(0.06)
Panel D: Effects by content domain					
Data	0.39	0.10***	0.11***	-0.01	0.01
	[0.21]	(0.02)	(0.02)	(0.02)	(0.02)
Geometry	0.50	-0.00	-0.14***	0.00	0.04***
	[0.29]	(0.02)	(0.02)	(0.02)	(0.01)
Number sense	0.61	-0.07***	-0.21***	-0.01	0.02
	[0.27]	(0.02)	(0.02)	(0.02)	(0.02)
Whole Number Ops	0.54	-0.03	-0.04**	-0.02	0.02
	[0.29]	(0.02)	(0.02)	(0.02)	(0.02)

Notes. This table provides descriptive statistics for the control group (column 1), control-group gains to midline (column 2), control-group gains to endline (column 3), the difference across treatment and control students at midline (column 4), and the difference across treatment and control students at endline (column 5). Outcomes in Panel A and C are standardized with respect to the control group at baseline. All other outcomes reflect proportions ([0,1]). Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s) and a vector of control variables selected via Lasso. * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE III
ITT effects on student learning, by program variant

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on main outcome									
Main outcome	0.08	0.13*	0.40***	-0.09	-0.02	-0.07	0.02	0.12*	-0.10
	[0.98]	(0.07)	(0.08)	(0.07)	(0.07)	(0.08)	(0.09)	(0.07)	(0.10)
Panel B: Effects on ASER test									
ASER (Baseline)>=1-digit	0.99	0.01	0.01***	-0.01**	-0.00	-0.00	-0.00	0.00	-0.00
	[0.09]	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ASER (Baseline)>=2-digit	0.90	0.06***	0.07***	0.00	-0.02*	0.02*	-0.00	-0.02**	0.02*
	[0.29]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ASER (Baseline)>=Subtraction	0.35	0.09***	0.24***	-0.02	-0.00	-0.02	-0.00	0.02	-0.02
	[0.48]	(0.02)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
ASER (Baseline)>=Division	0.10	0.01	0.11***	0.03	-0.02	0.04**	0.01	0.02	-0.01
	[0.30]	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)
Panel C: Effects by cognitive domain									
Higher-order	0.06	0.08	0.27***	-0.09	-0.01	-0.08	-0.03	0.08	-0.11
	[1.00]	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.09)	(0.07)	(0.11)
Lower-order	0.09	0.10	0.31***	-0.09	-0.00	-0.09	0.06	0.14**	-0.08
	[0.98]	(0.07)	(0.07)	(0.08)	(0.07)	(0.09)	(0.09)	(0.07)	(0.10)
Panel D: Effects by content domain									
Data	0.39	0.10***	0.11***	-0.01	-0.01	-0.00	0.00	0.02	-0.02
	[0.21]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)
Geometry	0.50	-0.00	-0.14***	-0.02	0.02	-0.03*	0.03	0.05***	-0.02
	[0.29]	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Number sense	0.61	-0.07***	-0.21***	-0.02	0.00	-0.02	0.00	0.03	-0.03
	[0.27]	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)
Whole Number Ops	0.54	-0.03	-0.04**	-0.03	-0.01	-0.02	0.00	0.03	-0.03
	[0.29]	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)

Notes. This table provides descriptive statistics for the control group (column 1), control-group gains to midline (column 2), control-group gains to endline (column 3), differences across experimental groups at midline (columns 4-6), and differences across experimental groups at endline (columns 7-9). Columns 4 and 7 compare the full program against the control group; Columns 5 and 8 compare the partial program against the control group; Columns 6 and 9 compare the two treatment variants against each other. Outcomes in Panel A and C are standardized with respect to the control group at baseline. All other outcomes reflect proportions ([0,1]). Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s) and a vector of control variables selected via Lasso. * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE IV
Heterogeneity in ITT effects on student learning

	(1)	(2)	(3)	(4)	(5)
Panel A: By gender					
Female	0.12	0.09	0.31***	-0.03	0.14*
	[0.98]	(0.07)	(0.08)	(0.06)	(0.08)
Male	-0.05	0.22**	0.51***	-0.10	-0.02
	[0.96]	(0.10)	(0.08)	(0.07)	(0.07)
Male vs Female	-0.13*	0.13**	0.20***	-0.07	-0.16**
	(0.07)	(0.06)	(0.06)	(0.05)	(0.07)
Panel B: By baseline writ level					
Bottom tercile	-1.06	0.66***	0.96***	-0.14	0.07
	[0.55]	(0.05)	(0.04)	(0.09)	(0.08)
Middle tercile	0.01	0.20***	0.41***	-0.06	0.04
	[0.25]	(0.04)	(0.04)	(0.07)	(0.08)
Top tercile	1.05	-0.31***	-0.10**	0.01	0.09
	[0.49]	(0.04)	(0.04)	(0.08)	(0.09)
Top vs bottom tercile	2.07	-0.97***	-1.06***	0.14	0.03
	(1.48)	(0.09)	(0.11)	(0.11)	(0.10)
Panel C: By district					
Bijapur	-0.04	0.19***	0.58***	-0.12	0.08
	[0.76]	(0.06)	(0.05)	(0.08)	(0.09)
Tumkur	0.19	0.12**	0.30***	0.03	0.06
	[0.04]	(0.05)	(0.04)	(0.09)	(0.10)
Tumkur vs Bijapur	0.22*	0.08	0.27**	0.14	-0.02
	(0.13)	(0.15)	(0.14)	(0.12)	(0.14)

Notes. This table provides descriptive statistics for the control group (column 1), control-group growth to midline (column 2), control-group growth to endline (column 3), the difference across treatment and control students at midline (column 4), and the difference across treatment and control students at endline (column 5). The outcome is students' overall math score, standardized with respect to the control group at baseline. Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s) and a vector of control variables selected via Lasso. * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE V

Heterogeneity in ITT effects on student learning, by program variant

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: By gender									
Female	0.12	0.09	0.31***	-0.08	0.02	-0.10	0.09	0.18**	-0.09
	[0.98]	(0.07)	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)	(0.09)	(0.12)
Male	-0.05	0.22**	0.51***	-0.14*	-0.06	-0.08	-0.08	0.04	-0.12
	[0.96]	(0.10)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.08)	(0.10)
Male vs Female	-0.13*	0.13**	0.20***	-0.06	-0.08	0.02	-0.17**	-0.15*	-0.02
	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
Panel B: By baseline learning level									
Bottom tercile	-1.06	0.66***	0.96***	-0.15	-0.11	-0.04	0.03	0.12	-0.09
	[0.55]	(0.05)	(0.04)	(0.11)	(0.10)	(0.11)	(0.11)	(0.10)	(0.13)
Middle tercile	0.01	0.20***	0.41***	-0.08	-0.03	-0.05	0.02	0.06	-0.04
	[0.25]	(0.04)	(0.04)	(0.09)	(0.09)	(0.10)	(0.11)	(0.08)	(0.11)
Top tercile	1.05	-0.31***	-0.10**	-0.07	0.09	-0.16	0.01	0.18*	-0.17
	[0.49]	(0.04)	(0.04)	(0.11)	(0.08)	(0.11)	(0.12)	(0.10)	(0.13)
Top vs bottom tercile	2.07	-0.97***	-1.06***	0.08	0.20*	-0.12	-0.02	0.06	-0.08
	(1.48)	(0.09)	(0.11)	(0.15)	(0.12)	(0.15)	(0.13)	(0.12)	(0.13)
Panel C: By district									
Bijapur	-0.04	0.19***	0.58***	-0.14	-0.09	-0.05	0.00	0.16	-0.16
	[0.76]	(0.06)	(0.05)	(0.10)	(0.10)	(0.12)	(0.12)	(0.11)	(0.15)
Tumkur	0.19	0.12**	0.30***	-0.06	0.09	-0.14	0.04	0.07	-0.03
	[0.04]	(0.05)	(0.04)	(0.10)	(0.12)	(0.13)	(0.12)	(0.12)	(0.13)
Tumkur vs Bijapur	0.22*	0.08	0.27**	0.08	0.18	-0.10	0.04	-0.09	0.13
	(0.13)	(0.15)	(0.14)	(0.14)	(0.16)	(0.18)	(0.17)	(0.17)	(0.21)

Notes. This table provides descriptive statistics for the control group (column 1), control-group gains to midline (column 2), control-group gains to endline (column 3), differences across experimental groups at midline (columns 4-6), and differences across experimental groups at endline (columns 7-9). Columns 4 and 7 compare the full program against the control group; Columns 5 and 8 compare the partial program against the control group; Columns 6 and 9 compare the two treatment variants against each other. The outcome is students' overall math score, standardized with respect to the control group at baseline. Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s) and a vector of control variables selected via Lasso. * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE VI
Robustness of results at endline

	Attrition			Sample definition				Outcome	Re-randomization
	IPW	Lee (lower)	Lee (upper)	Any BL	Any BL, written EL	No contamination	Complete strata	Written only	Rand. inference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Average effect	0.07	0.05	0.08	0.07	0.07	0.07	0.09	0.07	0.07
	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	[0.42]
Effect with events	0.01	0.01	0.02	0.02	0.02	0.02	0.04	0.02	0.02
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	[0.86]
Effect without events	0.12*	0.09	0.14**	0.12	0.12	0.11	0.13*	0.13*	0.12
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	[0.27]
Effect without events, among girls	0.18**	0.13	0.19**	0.18**	0.18**	0.16*	0.19**	0.18**	0.17
	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	[0.15]
Effect without events, among boys	0.04	0.03	0.07	0.05	0.04	0.03	0.07	0.06	0.05
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.08)	[0.64]

Notes. This table presents robustness checks for the paper's main findings at endline. Columns (1) to (3) investigates robustness to attrition. We present inverse probability-weighted estimates and Lee (2009) bounds. Columns (4) to (7) investigate robustness to alternative sample definitions. We present results for the study sample of students with any baseline test, the study sample of students with any baseline test and at least a written endline test (but not necessarily an oral endline test), a sample where we remove a randomization stratum with contamination (one treatment school of the group without events received an event), and a sample of strata where no schools had to be dropped after baseline (two schools had zero attendance at baseline). The outcome is students' overall math score, standardized with respect to the control group at baseline, except in column (8). Column (8) investigates robustness to measurement. We present results for an outcome measure that uses written test items only, ignoring students' performance on the oral test. Standard errors in parentheses (clustered at the Panchayat level). Column (9) investigates robustness to the re-randomization procedure used for treatment assignment. We present randomization inference (RI)-based p-values (in brackets), where we repeated the same re-randomization procedure in 5,000 RI iterations. All estimations include randomization strata fixed effects (F.E.s) and a vector of control variables selected via Lasso. * significant at 10%; ** significant at 5%; *** significant at 1%.