Large-Scale Efforts to Improve Teaching and Child Learning: Experimental Evidence from India*

Andreas de Barros†
Johanna Fajardo-Gonzalez‡
Paul Glewwe§
Ashwini Sankar¶

November 7, 2022

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 program's combined effect had at most small, statistically insignificant impacts on learning. School value-added models suggest a negative relationship between activity-based instruction and test score gains. The addition of contests worsened instructional quality (especially classroom culture), and we can rule out that the contests added even small improvements in learning.

*This study has been conditionally accepted at the Journal of Development Economics (JDE), through its pre-results review process. The authors thank the editor of that journal 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, Andrea Guariso, and Tavneet Suri provided additional helpful 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 Jack Cavanagh for an expertly conducted code replication. 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 views expressed in this paper are those of the authors and do not necessarily represent those of the United Nations, including UNDP, or the UN Member States. The authors have no conflicting interests to declare.

†Postdoctoral Associate, Department of Economics, MIT. E-mail: debarros@mit.edu.
‡Economist, United Nations Development Programme. E-mail: fajar016@umn.edu
§Professor, Department of Applied Economics, University of Minnesota. E-mail: pglewwe@umn.edu.
¶Senior Researcher, Health Policy Division, Minnesota Department of Health. E-mail: sanka010@umn.edu.

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 training designed to improve instructional quality, and (2) provision of teaching materials and teacher training *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 more resources, and training on how to use them, is expected to 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 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 strategy 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 typically used in Indian schools. It thus aligns with India's National Education Policy 2020, which focuses on foundational mathematics skills and promotes activity-based, experiential learning in primary grades. The program also conducts community-led contests that convene stakeholders to witness the mathematical performance of school children during a public assessment. It is a collaboration between the state government and an Indian non-governmental organization 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.

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 the 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 students' learning of mathematics. Specifically,

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

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. The variant without contests raised student learning by 0.12 SDs ($p = 0.11$). In contrast, the estimate for the variant with community contests is almost zero (0.01 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 variants raised student attitudes toward mathematics (0.09 SDs). Both also succeeded in promoting pre-specified dimensions of teaching practices that are expected to promote students' critical thinking, autonomy, and collaboration (0.16 to 0.27 SDs). However, adding community contests to the intervention led to sizeable negative effects on students' classroom environment in the study periods after the community contests had been conducted (-0.40 to -0.49 SDs). Parent 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 ($p < 0.10$) 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, although one robustness check (employing randomization inference) yields a p-value somewhat above the 10-percent critical value ($p = 0.14$) for girls' math scores. Yet, our finding that the addition of community contests did not lead to higher impacts holds for all robustness checks.

Finally, we present two ancillary analyses. First, we compare our estimates with those of a systematic review of other rigorous studies that investigated the same pedagogical approach. We show that, while expert opinion strongly recommends this particular teaching practice, our findings can rule out the positive effects found in nearly all (26 out of 28) prior studies. Second, going beyond the experimental design of our study, we calculate school value-added models, both in terms of schools' effectiveness in raising student learning and their effectiveness in improving student attitudes toward mathematics. We find that the teaching practices that are promoted by the program are negatively correlated with school effects on test scores (and we do not find an association with school effects on student attitudes). We also show that these two types of school effects are orthogonal to each other.

This paper makes three contributions to the literature. The first is to the broader literature on interventions that seek to improve teacher effectiveness in less-developed countries. 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 (Evans and Popova 2016; Popova et al. 2022).[2] Successful on-site training and teacher coaching programs also support this finding (Cilliers 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 2022), and training is provided remotely (Cilliers et al. 2022). In fact, only three of the teacher training studies reviewed by Schaffner, Glewwe, and Sharma (2021) examined programs that were implemented in more than 300 schools—and all three failed to improve student learning.[3]

The article also contributes to the economics of education literature on value-added models. While these models are increasingly common for developed countries, only a handful of published studies have employed them for less-developed countries (Andrabi et al. 2011; Araujo

2. In contrast, standalone, off-site training events that target teachers' general skills are commonly not related to changes in teaching practices or improvements in student learning. For example, the Global Education Evidence Advisory Panel refers to in-service, general-skills, off-site teacher training as "usually not a good investment", with "little evidence showing that the typical stand-alone general-skills in-service training is cost-effective" (The World Bank 2020, 21).

3. Loyalka et al. (2019) examined a national government-run teacher training program in China that operated at the secondary level (grades 7-9), and found that it was ineffective at raising students' test scores. Berry et al. (2020) evaluated a national program in India, as implemented in the state of Haryana, that trained teachers to use "continuous and comprehensive evaluation" in order to better monitor their students' progress and so adjust their teaching to match their students' levels of learning; this program, which focused on grades 2-5, also failed to increase students' test scores. Finally, Schaffner, Glewwe, and Sharma (2021) examined a national government-operated teaching training program in Nepal that provided math and science teachers in grades nine and ten with new pedagogical methods to teach those subjects, and this program also failed to raise student learning. For a review of effect sizes in international education studies, see Evans and Yuan (2022). The only at-scale program that raised student learning, Majerowicz and Montero (2018), was not reviewed by Schaffner, Glewwe, and Sharma (2021).

et al. 2016; Bau and Das 2020; Singh 2015). Whether in developed or developing countries, even fewer studies have been able to relate schools' or teachers' value-added to detailed classroom observations of teaching practices (Araujo et al. 2016; Blazar and Kraft 2017). Our ancillary analyses add to evidence that one such practice may reduce student learning (Berlinski and Busso 2017), and we rule out the positive effects found by a large body of related efficacy trials from the U.S. (Fuchs et al. 2021). Our study also adds to a nascent value-added literature that examines the multidimensionality of educational effects on assessment-based vs. non-test outcomes (Beuermann et al. 2022; Blazar and Kraft 2017; Jackson 2018).

Our third 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.[4] 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) found no effects for a similar intervention that was 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 failed to raise student 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 impacts of education interventions if parents disagree with novel pedagogical practices (Wolf et al. 2019).

4. A related strand of literature examines information-provision campaigns that seek to remove information frictions in educational markets; a recent World Bank (October 2020) report cites such campaigns as "great buys". For an example from Pakistan, see Andrabi, Das, and Khwaja (2017); for an example from Chile, see Mizala and Urquiola (2013). While our study focuses on potential complementarities across and demand- and supply-side factors, yet another related strand of literature considers complementarities among supply-side factors. For an example from Tanzania, see Mbiti et al. (2019); for an example from Colombia, see Barrera-Osorio et al. (2018).

I. Research Design

I.A. Context

From ages 6 to 14, schooling in India is free and compulsory. Elementary education runs from grades 1-8, of which grades 1-5 are "primary" education (the focus of this study) and grades 6-8 are "upper primary". In 2018, India had 1,255,841 schools serving "primary" grades, of which 69 percent (860,790) were managed by state and local governments (NIEPA 2018).

In 2020, India passed its new National Education Policy (NEP), which codified the country's shift in focus toward foundational numeracy and literacy learning in the primary grades. In terms of mathematics pedagogy, NEP and a subsequent national mission (National Initiative for Proficiency in Reading with Understanding and Numeracy, "NIPUN Bharat") identified "joyful and experiential learning" as a focus area, advising that classroom interactions include "toys, games, [...], etc. to be used extensively for teaching through play/discovery/game/art/activity-based pedagogy" (Ministry of Education 2021, 201). India now promotes such "experiential learning" both nationally (CBSE 2022) and through state initiatives. For example, new programs in Uttar Pradesh and Madhya Pradesh aim for students to learn concrete mathematical concepts first, before moving to abstract understanding; we investigate a similar state-wide initiative.

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 before scaling up to the entire country. First, the state is large, ranking sixth in terms of area and eighth in population (MHA 2012). Second, it exemplifies how high enrollment and additional inputs may not raise student learning. It has very high enrollment (over 99 percent of rural children ages 5-14 are in school), attendance (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, primary students' arithmetic skills rank Karnataka near the bottom of India's states (e.g., less than 20 percent of rural government-school students in grade five can do basic division) (ASER 2018). Third, other states often mimic Karnataka's education policies; for example, Odisha recently adopted the program evaluated in this paper.

I.B. Intervention

We conducted this study cooperating with the Akshara Foundation, a large NGO dedicated to ensure quality pre-school and primary education in India. Founded in the year 2000, the organization works with several state governments to support primary education in government-led schools. The Akshara Foundation's Ganitha Kalika Andolana (GKA) intervention combines provision of new instructional materials, related teacher training, and community engagement to improve primary-school students' mathematics abilities. This subsection summarizes the program's two main components (see Appendix B for more detail on the intervention).

The program was started in 2011 for 249 government primary schools in Bangalore Rural District. Karnataka's Government has since committed to scale it up to all of the state's 44,000 Government primary schools, in a phased manner. In 2017, another Indian state, Odisha, began to implement GKA, expanding it to about 30,000 schools by 2020. We conducted the study during the state-wide scale up in Karnataka; we did not alter the intervention.

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

The program's first component provides 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 that follows a "concrete-representational-abstract" (CRA) model.[5] The TLM kits include items such as an abacus, a set of shapes, and measuring kits. Each item maps into mathematical concepts required by the state curriculum.

Expert teachers provide training to the primary school teachers. These off-site training sessions are held during the state's scheduled teacher training, replacing its content; they are not "additional" training sessions, which keeps costs neutral. The training focuses on enabling teachers to create activities using the TLM kits. After this initial training, a block-level field

5. Under the "concrete-representational-abstract" (CRA) model, students: first, develop conceptual understanding by manipulating objects; next, learn how pictures, numbers, and symbols represent objects; and last, master mathematical problems using only abstract numbers and symbols. CRA is loosely based on a learning theory with three "Stages of Representation": enactive, iconic, and symbolic (Bruner and Kenney 1965). CRA is sometimes interchangeably referred to as the "Concrete, Pictorial, Abstract" (CPA) approach.

coordinator supports the teachers as they implement this new teaching method.

I.B.2. Community contests

The second component is the community contests. These Gram Panchayat Mathematics Contests ("GP contests") convene stakeholders to observe the mathematical performance of students. They are intended to encourage parent engagement and community participation, which can put pressure on teachers to improve their teaching, thereby raising student learning.

Contests start with a math test for students from any government primary school in the GP. After the test, participants discuss the GKA program and other education issues, focusing on students' learning and the quality of instruction. Next, the assessment results are announced, the top three students are recognized, and other education performance statistics are presented to community members. Following the contest, a letter is sent to local leaders and to the school's School Development and Monitoring Committee (SDMC), summarizing test scores for each participating school. GPs are free to decide whether they would like to hold a contest and, while the Akshara Foundation initiates these contests in participating GPs, the GP and other local sources pay for all operational expenses.[6] In any given school year, a GP holds at most one contest.

I.B.3. Program costs

We estimate that across the two program versions, the average program cost is USD 7.4 per student. The variant of the program without GP contests has a cost of USD 6.8 per student; the variant with the GP contests is USD 8 per student. The program's cost thus falls around the middle of the costs of interventions that have improved student learning in India's public schools.[7]

6. Students who are present cannot opt out of participating in the contests, but they may be absent on the day of the contest. Parents are free to decide whether they would like to attend a contest.

7. One example is a study on remedial education, for which math and language learning increased at a cost of 4.5 USD per student (Banerjee et al. 2007). Another example is Duflo, Hanna, and Ryan (2012), who find that a program that provided teachers with incentives to work improved student learning in math and language at a cost of USD 7.5 per student. More recently, Nyqvist and Guariso (2021) document how the combination of targeted instruction with out-of-school study groups increased math test scores at a cost of USD 17.2 per student.

I.C. Sampling and sample

We implemented the study in two of Karnataka's 30 districts: Tumkur and Vijayapura. We selected these two districts to maximize the study's geographic spread and representativeness within the state.[8] We considered only "Higher Primary Schools", which end in grades 7 or 8, and prior to sampling, we excluded schools where the medium of instruction was not Kannada and schools with fewer than five students in grade four in the previous school year (including those schools that did not teach grade four at all). We also excluded Gram Panchayats (GPs) with fewer than three eligible schools. We first randomly sampled 98 GPs from these two districts. Within each GP, we randomly sampled three schools, yielding 294 schools. Two schools were removed, reducing our sample to 292 schools, after we discovered that they had no fourth-grade students; this removal occurred prior to randomization into treatment and control schools.

The sampling strategy ensured that half of these GPs and schools were drawn from each of the two districts. In each district, we randomly selected 49 GPs using "probability-proportional-to size" (PPS) sampling. Next, we randomly selected 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 Figures A1 and A2 show the studies two districts and the randomly selected GPs and schools.

At baseline, 5,227 fourth-grade students were enrolled in the study's 292 schools, of which 4,026 (77.0 percent) were present for the baseline data collection. This number is similar to other large-scale studies in India; for example, Goodnight and Bobde (2018) report a 73.1 percent attendance rate for India's government primary schools.

These 4,026 students are the study's sample. At baseline, they were, on average, 9 years and 2 months old. About 53 percent were female.[9] Of these students, 3,971 (98.6 percent) took the written baseline test, and 3,881 (96.4 percent) took the written *and* oral baseline tests (described in Section III). We focus on students with both tests, but we also present robustness checks for the sample with only the written test and the full sample (with or without baseline tests).

To analyze intermediate outcomes, we interviewed sub-samples of students and parents by

8. In Section IV., our results suggest that these districts indeed showed different learning levels at baseline; control-group students in these districts also differed in terms of their progress from baseline to endline.

9. These students' age and gender numbers are approximate, as they are missing for 2.0 percent of the students.

randomly selecting (up to) eight students and parents, per school. We drew new samples of students and parents for each survey round. More specifically, we conducted 1,924 student surveys in the first process monitoring round, 1,875 in the second round, and 1,861 in the fourth round (we did not conduct student surveys in the third round). We conducted 1,967 parent interviews in round three (we did not conduct parent surveys in the remaining rounds).

I.D. Randomization

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. Within each district, we used baseline test scores on the one-on-one test (described below) to create quadruplets of GPs with similar academic performance. For each stratum of four GPs, two were randomly selected to participate in the GKA program, leaving the other two as "controls." Thus, 49 GPs and their selected schools were assigned to receive the program; the other 49 and their selected schools were "controls."[10] We repeated this randomization procedure ten times and selected the one with the greatest balance (see Appendix D for details).

Finally, we randomized all 49 treatment GPs into two arms: one of 24 GPs with community contests, and one of 25 GPs without those contests. Both treatment arms received the kits and related training. Figure I depicts the study schools by treatment status. In Section IV.A.1. below, we check whether the randomization led to comparable groups.

II. Hypotheses, Outcomes, and Data

II.A. Primary hypothesis and main outcome of interest

Our primary hypothesis is that the program improves students' mathematics learning. We measure this using students' scores on (1) standardized math tests; and (2) a one-on-one test of basic math skills.

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 ad-

10. 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).

ministered to students in groups.[11] Assessments had 30-35 multiple-choice items and students had a one-hour time limit; students typically took about 45 minutes to complete each test.

Test items are mapped to the official state curriculum, and include items one or two years below grade level. They have been administered in similar contexts in India for large-scale assessments. None of them is from Akshara Foundation's internal item bank. From the previous administrations, we used item response theory (IRT)-based item characteristics to maximize the assessments' test information. See Jacob and Rothstein (2016) for an introduction to IRT.

Due to its salience in India, we also administered the "ASER" test of basic arithmetic skill (cf. ASER 2018)[12] to the full sample of students. 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 ordered 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.

We estimate each student's ability using a two-parameter logistic (2PL) IRT model (Birnbaum 1968). We used anchor items across test rounds (baseline, midline, endline) to link all rounds onto a common, continuous ability scale (Kolen and Brennan 2004).[13]

We describe in more detail the test design and related validity evidence in Appendix C. That evidence confirms that the tests did not display floor or ceiling effects. It also indicates that our test items discriminate well for student ability and that the tests exhibit low levels of noise.

II.B. Secondary hypotheses and related outcomes

We pre-specified three sets of secondary hypotheses. We describe their respective outcomes here.

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

12. 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.

13. Our registered report did not discuss how to combine oral and written items. 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. This 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.

Measures of sub-competencies. First, we investigate the program's impact on student learning along more fine-grained sub-competencies of mathematical skill, which can be grouped into two sets of (more fine-grained) domains: content domains and cognitive domains. The tests capture students' ability in four content domains—number sense, whole number operations, shapes and geometry, and data display, measurement, and statistics—and two cognitive domains—knowing, and reasoning and applying.

Intermediate outcomes. Our second set of secondary hypotheses focuses on three types of intermediate outcomes. The first is teachers' instructional behaviors. After the program's implementation, we used unannounced classroom observation visits to measure instructional quality, time-on-task, and instructional behaviors in treatment and control schools. These visits were scheduled to follow the study's sample of students—not a given mathematics teacher—so we focused on the instruction these students actually received, regardless of whether their teachers changed over time. We conducted one round of these visits 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*". It focuses on three broad domains of instructional quality—Classroom Culture, Instruction, and Socio-emotional Skills—as well as nine narrower sub-domains. We pre-specified that we would expect to find impacts on three of the nine sub-domains: teaching practices related to critical thinking, autonomy, and social and collaborative skills.[14] In addition, we complement *Teach* with two ancillary data sources for teachers' instructional behaviors: Teacher surveys (during school visits) and surveys of sub-samples of students.

The second type of intermediate outcomes concerns parental involvement and community engagement. The student interviews included questions on parental involvement in their child's math education. The teacher interviews elicited teachers' perceptions of parental involvement, including when they last communicated with a parent. Interviews with the sub-sample of parents asked about their involvement in their child's math education. To measure community

14. Appendix Figure A3 provides mean *Teach* scores, for the control group.

engagement, we asked headmasters about the activities of their schools' School Development and Monitoring Committee (SDMC); these committees formalize community involvement in school management and school improvement efforts. We also asked headmasters about parents' meetings with teachers.

The third type of intermediate outcomes is student attitudes toward mathematics. We used surveys of the sub-sample of students to measure their attitudes toward mathematics learning. 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.

Implementation fidelity. The third set of secondary hypotheses examines implementation fidelity in treatment schools.[15] We organize primary and secondary data on implementation fidelity by the program's two main components: 1. Teacher training and additional teaching inputs; and 2. Community contests ("GP contests").

Regarding the former, the Akshara Foundation provided us with administrative records on teachers' participation in GKA training sessions. We augmented these data by asking teachers, in our teacher survey, 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 by Akshara staff.

Turning to the latter, the research team attended all community contests ("GP contests"). During the contests, we recorded individual student attendance, including unique student IDs. At each contest, the research team also recorded parents' attendance. The student survey questionnaire also asked the students whether they had participated in the GP contests.

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

II.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.

II.D. *Additional data and timeline*

In addition to measuring students' skills, we collected their demographic information (including gender, date of birth, and parents' name) to use as additional covariates and to facilitate tracking of them over the study's multiple rounds of data collection. 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, using GIS software to match each school's location to its respective village.

Appendix Figure A4 depicts the study's timeline, for both program implementation and data collection. The data collection began with the November 2018 baseline survey, followed by four rounds of process monitoring in 2019 (February, August, November, and December) and midline (September 2019) and endline (February 2020) assessments.[16] All students started the study in grade four, and a new school year began after the first round of process monitoring. We tracked individual students irrespective of whether they moved to grade five (almost all did), yet focused process monitoring rounds two to four on grade-five classrooms.

16. We used J-PAL'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).

III. Empirical Strategy

III.A. Statistical model

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

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

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} represents test scores. In our secondary analyses, Y_{isgr}^{t} is either measures of sub-competencies or potential mediating variables. The α_r terms are strata fixed effects, T_{gr} is the treatment dummy and ϵ_{isgr}^{t} is the residual. 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, and $\boldsymbol{X}_{isgr}^{t=0}$ is a vector of baseline controls selected by a LASSO procedure on student age, gender, school-level DISE data, and village-level census data. The coefficient of interest, β^t, is the program's intent-to-treat (ITT) effect, for follow-up round t.

Our secondary analyses estimate the additional effect of community contests, specified as:

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

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 with or without 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).

We also use specifications that allow for heterogeneous treatment effects, by interacting potential moderators with the treatment indicator. We do this using the following specification:

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

Here, $F_{isgr}^{t=0}$ is a moderating variable of interest (e.g., student gender); all else is as defined above.

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

III.B. Statistical methods

We estimate OLS regressions. For the ASER data, we create binary outcomes, so we estimate linear probability models. We cluster standard errors at the GP level (cf. Abadie et al. 2022).

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).

We describe the statistical methods in more detail in Appendix D.

IV. Results

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

IV.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 84 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 28 percent for the control group, but this is reduced to 19 percent from baseline to endline. Although these attrition rates seem high, we find comparable rates in other studies. In a randomized experiment of Salvadorian migrants to estimate the impact of an educational subsidy on the demand for education (Ambler, Aycinena, and Yang 2015), the attrition rate was 27% in the one-year follow-up. In another randomized experiment studying students' behaviors in a French middle school (Avvisati et al. 2014), the one-year follow-up had an attrition rate of 17%. In studying neighborhood peer effects on children's school enrollment decisions, Bobonis and Finan (2009) found an attrition rate of 20%. Attrition from baseline to midline in our study 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.[17]

IV.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 A4, 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 assessment, we estimate an exposure of 19 weeks. The exposure until endline was 37 weeks.[18] Our calculations, based on the official school calendar but removing any days with school closures (e.g., due to local festivals and holidays, or due to floods), indicate that the effective number of days that schools were open over the study period was 215 days.

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. This is not surprising because

17. As shown by Ghanem, Hirshleifer, and Ortiz-Becerra (2020), differential attrition is not a source of concern regarding internal validity—selective attrition is. Following Ghanem, Hirshleifer, and Ortiz-Becerra (2020), we conduct a formal test using the study's baseline learning levels. While attritors differed from non-attritors ($p < 0.01$ at midline and endline), we do not find evidence of selective attrition across experimental groups ($p = 0.34$ at midline; $p = 0.55$ at endline; these results are available from the authors upon request). This corroborates that our study's findings are internally valid for the analytical sample.

18. From the random assignment of treatment GPs to community contests (in July 2019), the exposure until endline was approximately 50 percent shorter (the median interval between a contest and the endline assessment was five months). It is not uncommon for education studies to observe short-term effects over a similar time span. For examples, see Li et al. (2014), Duflo, Hanna, and Ryan (2012), and Banerjee et al. (2007).

the 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. 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, and no teachers in control 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 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) treatment-group teachers reported having received the GKA kit. More teachers in treated schools (38 percent) conduct group activities on a daily basis than teachers in control schools (15 percent.) Most (60 percent) treatment-group teachers reported using the GKA kit for math classes in 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 much 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 last indicators use GP contest data and reveal that only 1.4 percent (14 out of 1,018) of parents attended the GP contests,[19] although 73 percent of students participated in the

19. We also conducted a parent survey at endline. Eleven percent of parents (self-)reported that they had attended a math contest. Our preferred data source relies on our research team's observations during the contests.

contests.

IV.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 ($p < 0.10$). 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 ($p > 0.10$)—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.

IV.C. *Secondary results*

IV.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 34 percent. At endline, improvements are more pronounced and they also include an 11 percentage-point increase in the proportion of students who know division (from a base level of 9 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") versus 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 ($p > 0.1$).

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).[20] 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 students' 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.

IV.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 type without community contests (0.12 SDs, $p = 0.11$). 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.

IV.C.3. Heterogeneous effects

We now investigate whether the effects on student math learning differ for three different (pre-specified) 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.

20. 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, those gains can be readily interpreted.

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 ($p < 0.10$), 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 ($p < 0.10$). 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.

IV.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.[21] 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 ($p > 0.10$). This null finding for the overall index masks a positive impact of 0.27 SDs on the pre-specified dimension of teaching behaviors (e.g., whether the teacher promotes student autonomy), 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 ($p > 0.10$). This overall finding is driven by a 0.16 SD improvement for the pre-specified dimension of teaching quality ($p < 0.05$) and a 0.17 SD improvement for the subdimension related to socioemotional skills ($p < 0.05$). We document a (statistically insignificant) 0.08 SD improvement in instructional quality. Once community

21. 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.

contests are removed from the program, the coefficient for the dimension of classroom culture is close to zero and statistically insignificant.

Recall that the randomization of treatment GPs to contests occurred after data collection Round 1, and the launch of contests approximately coincided with data collection in Round 2; only two treatment Gram Panchayats (and their six schools) received the contest prior to Round 2 data collection (compare to Appendix Figure A4). This allows us to observe the causal effect of adding the community contests to the intervention, across the study period: in Rounds 1 and 2, we would not expect any differences across the two treatment groups, but effects may materialize after Round 2. As shown in Table VI, the negative effects of adding community contests to the intervention coincide with this timeline. 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). As shown in the table's even-numbered columns, these results hold when we add Round 1's school-level *Teach* scores as a covariate in the estimation of effects in Rounds 3 and 4. Appendix Table A2 does the same for each of the nine underlying *Teach* indicators.

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. 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 do not differ in comparison to control schools.

In the bottoms of panels (a) and (b) 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.

IV.D. Robustness checks

As discussed earlier, we subject the study's main findings to a series of robustness checks. Table VII 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, and 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 above the program variant without the contests.

V. Ancillary Analyses

V.A. Connection to other education studies

Null findings (and findings that can rule out substantial impacts) can be particularly informative if they challenge expert opinion. To assess whether this is the case for our paper, we compare our findings with those from a recent systematic review of rigorous education studies on how to assist primary school students who struggle with mathematics (Fuchs et al. 2021). This review strongly supports the pedagogical approach promoted by India's GKA program, whereby students are prompted to move from concrete and semi-concrete to abstract representations of mathematical concepts. From their assessment of 28 studies, the review's expert panel found that there is strong evidence in support of this pedagogical approach. They concluded that there is a preponderance of evidence of positive effects, with strong external validity, which provides a high degree of confidence that this pedagogical practice is effective.

In Figure V, we compare the findings from the 28 studies identified by this systematic review to our findings in this paper (Appendix Figure A5 provides the same information by program variant). The panel on the left plots effect sizes against students' exposure to an intervention (in weeks), while the panel on the right plots effect sizes against each study's sample size. This comparison suggests that all but two of the previously reported effects are outside the confidence intervals that we find for the treatment effects of the GKA program (at midline and endline, respectively). This occurs even though, at endline, this intervention had provided the longest exposure duration to students. The comparison also shows that previous studies predominantly relied on small-sample efficacy trials (with a median sample size of 180 students). Such small sample sizes will yield significant results only for large impacts. If there is a tendency for insignificant results not to be published, then this could explain why these 28 studies found large impacts: perhaps many other studies found smaller impacts, but were never published.

V.B. Associations between teaching practices, student learning, and student attitudes

Teaching is multidimensional and, in developed countries, some teaching practices that improve students' performance on written assessments have been found to be unrelated to, or even negatively related to, other non-test outcomes (Beuermann et al. 2022; Blazar and Kraft 2017; Jackson 2018). Here, we explore whether schools' ability to improve test scores is correlated with the particular pedagogical approach promoted by the GKA program. We also explore whether schools' ability to increase test scores is correlated with their ability to improve students' attitudes toward mathematics. These analyses are exploratory, going beyond our registered report, and beyond the study's experimental design. Yet, we believe that they suggest an explanation for why the program did not lead to substantive impacts on student learning, despite its positive effects on a pre-specified dimension of teaching practices and its positive effects on student attitudes towards mathematics.

We begin by estimating value-added models of schools' ability to improve test and non-test outcomes. We estimate the following school fixed effects model

$$Y_{isg} = \alpha_{sg} + \delta' \boldsymbol{X}_{isg} + \epsilon_{isg} \tag{4}$$

where Y_{isg} is an outcome measure (test scores at endline or attitudes towards mathematics) for student i in school s and Gram Panchayat g, α_{sg} is a school fixed effect, $\boldsymbol{X}_{isg}$ is a vector of covariates measured at baseline, including baseline test scores, and ϵ_{isg} is an i.i.d. error term.

To account for the potential sorting of students into schools, other studies in the value-added literature usually estimate *classroom* fixed effects relative to the school average (e.g., Araujo et al. 2016; Bau and Das 2020; Chetty et al. 2011). In our context, schools usually have only one classroom and one teacher per grade; therefore, we can estimate only *school* fixed effects relative to the Gram Panchayat average (accounting for the potential sorting of students into

Gram Panchayats). This demeaned school effect, denoted by λ_{sg}, is

$$\lambda_{sg} = \alpha_{sg} - \frac{\sum_{s=1}^{S_g} N_{sg}\alpha_{sg}}{\sum_{s=1}^{S_g} N_{sg}} \tag{5}$$

where S_g is the number of schools in a Gram Panchayat (usually three schools), and N_{sg} is the number of students in school s in Gram Panchayat g.

Let (λ_{sg}) be the contribution of (variation in) school quality to (variation in) student-level outcomes. Then, $V(\hat{\lambda}_{sg})$ may overestimate $V(\lambda_{sg})$ due to sampling error. Therefore, following Chetty et al. (2011), we apply the following shrinkage procedure

$$V\left(\lambda_{sg}\right) = V\left(\hat{\lambda}_{sg}\right) - E\left\{\frac{\left(\sum_{d=1}^{S_g} N_{dg}\right) - N_{sg}}{N_{sg}\left(\sum_{d=1}^{S_g} N_{dg}\right)}\sigma^2\right\} \tag{6}$$

where σ^2 is the within-school variance of the student-level residual ϵ_{isg}.[22]

Our results point to substantial differences in schools' ability to boost test scores. We find that, corrected for sampling error, a one-standard-deviation increase in school quality is associated with a 0.37 standard-deviation increase in student learning. In contrast, schools' value-added for the non-test outcome is comparatively smaller; we find that a one-standard-deviation increase in school quality is associated with a 0.19 standard-deviation increase in student attitudes towards mathematics.

Next, we relate λ_{sg} to teachers and teaching practices (focusing on the pre-specified subdimension of practices associated with the GKA program). Table VIII shows the results of regressing the estimates of λ_{sg} for student test scores on a vector of teacher characteristics, the overall *Teach* index, and the subdimensions of teaching practices.[23] Column (1) shows observable teacher characteristics can explain only a very small fraction of the variance of student learning. Column (3) shows that the pre-specified measure of teaching practices is *negatively* related to student learning (and none of the other subdimensions is distinguishable from zero, at

22. Note that our measure of school value-added includes school effects, teacher effects, and idiosyncratic classroom shocks. We observe only one cohort of students and usually do not observe multiple classrooms per school; thus we cannot disentangle these effects from each other (as in Araujo et al. (2016) or Bau and Das (2020)).

23. Following Bau and Das (2020), in these regressions that include value-added measures on the left-hand side, we do not use a shrinkage correction.

conventional significance levels). Appendix Table A3 repeats this analysis for student attitudes. We do not find a significant relationship between the pre-specified dimension of teaching and attitudes; however, a more welcoming classroom climate appears to be positively related to student attitudes toward mathematics.

Thereafter, we investigate the correlation between the estimates of λ_{sg} for student test scores and λ_{sg} for student attitudes. In Appendix Figure A6 we show that these two dimensions are unrelated to each other, with a correlation coefficient of 0.03 ($p = 0.56$).[24]

Finally, recall that we cannot account for the systematic sorting of better-performing students to particular schools within Panchayats, nor for the adoption of better teaching practices in classrooms with higher baseline scores. Appendix Table A4 suggests that this phenomenon exists. Yet, note that better baseline scores correlate with better instruction and greater adoption of those teaching practices that are related to the subdimensions of teaching practices. We expect that these schools should be *more* productive than others and, therefore, we interpret our finding of a negative relationship between school value-added and activity-based instruction as a conservative estimate of (the absolute value of) the true association.

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 educational interventions to increase student learning. In India, both

24. Observationally, students' test scores are correlated with their attitudes, however. In the control group, students with a one-standard-deviation better attitude towards mathematics performed 0.21 standard deviations higher on the endline test ($p < 0.01$).

national and state-wide initiatives have started to promote one particular pedagogical approach: Activity-based instruction of mathematics, following a "concrete-to-abstract" model. Based on a large body of efficacy trials from developed countries, expert opinion concludes that this approach is highly effective. Yet, such interventions have rarely been evaluated at scale.

Our research investigates a related, large-scale effort to improve teaching and student learning. 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 the 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 randomly assigned the treatment group into two sets of Gram Panchayats. One set received the full program 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-four 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-four students (most of whom moved to grade five during the study period), 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 students' learning of mathematics. Analysis by gender finds a marginally significant impact of 0.14 standard deviations ($p < 0.10$) for girls' math scores, but no effect for boys. Even so, our estimates can rule out almost all positive estimates from previous efficacy trials, which led expert opinion to strongly recommend the program's pedagogical approach.

We also find a lack of impact for the addition of community contests, which is surprising and potentially important for policy. Both types of interventions led to positive impacts on a pre-specified dimension of instructional quality (0.16-0.27 SDs) and on teaching practices that promote socioemotional skills (0.17-0.24SDs). We also find marginally significant, positive effects of the intervention variant without community contests on grade 4 students' learning of mathematics (0.12 SDs, $p = 0.11$). Yet, the estimate for the program variant with contests is close to zero and we can rule out with precision that the addition of contests led to sizeable learning gains (added effects of 0.07 SDs or more are ruled out at 95 percent confidence).

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, instructional quality across the two treatment groups was indistinguishable, but after the contests, there are large negative effects on overall instructional quality 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 (-0.40 to -0.49 SDs, in the remaining study period after the contests had been conducted). These negative impacts of the contests suggest that efforts to raise community engagement may "backfire", triggering undesired behaviors in public schools.

Finally, our ancillary analysis of value-added models suggests that the schools which had greater adherence to the activity-based teaching practices promoted by the program exhibited *lower* productivity in raising student test scores. This (non-experimental) finding may explain why the program's positive impact on instructional quality, and especially on this dimension of instructional practices, did not coincide with substantive improvements in test scores. Hence, as governments try to promote activity-based learning at scale, they cannot expect that these efforts will necessarily lead to test score gains; rather, such efforts may even come at the expense of test score gains. At the same time, we also document how schools' productivity in terms of student performance on assessments was orthogonal to their ability to improve student attitudes towards mathematics. This suggests that the effects of activity-based instruction may be multi-dimensional. Unfortunately, through this study, we cannot rule out that the given pedagogical approach has desirable effects on other, non-test dimensions of student learning, which we did not measure.

Future research could go in several directions. First, program impacts may 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, and on socio-emotional skills particularly, 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 likely be very valuable. Finally, the negative impact of the community contests on classroom culture implies that further research is needed on what types of interventions, when applied to public schools, may lead to unintended negative consequences.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* (October): qjac038.

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.

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.

Ambler, Kate, Diego Aycinena, and Dean Yang. 2015. "Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador." *American Economic Journal: Applied Economics* 7, no. 2 (April): 207–32.

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.

Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc. 2011. "Do Value-added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3 (3): 29–54.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107 (6): 1535–1563.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131 (3): 1415–1453.

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

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.

Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin. 2014. "Getting Parents Involved: A Field Experiment in Deprived Schools." *Review of Economic Studies* 81 (1): 57–83.

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.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–1264.

Barrera-Osorio, Felipe, Sandra García, Catherine Rodríguez, Fabio Sánchez, and Mateo Arbeláez. 2018. "Concentrating Efforts on Low-Performing Schools: Impact Estimates from a Quasi-Experimental Design." *Economics of Education Review* 66:73–91.

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.

Bau, Natalie, and Jishnu Das. 2020. "Teacher Value Added in a Low-Income Country." *American Economic Journal: Economic Policy* 12, no. 1 (February): 62–96.

Berlinski, Samuel, and Matias Busso. 2017. "Challenges in educational reform: An experiment on active learning in mathematics." *Economics Letters* 156 (July): 172–175.

Berry, James, Harini Kannan, Shobhini Mukherji, and Marc Shotland. 2020. "Failure of frequent assessment: An evaluation of Indias continuous and comprehensive evaluation program." *Journal of Development Economics* 143 (March): 102406.

Beuermann, Diether W, C Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. 2022. "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output." *The Review of Economic Studies* (June): rdac025.

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.

Blazar, David, and Matthew A. Kraft. 2017. "Teacher and Teaching Effects on Students Attitudes and Behaviors." *Educational Evaluation and Policy Analysis* 39, no. 1 (March): 146–170.

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.

Bobonis, Gustavo, and Frederico Finan. 2009. "Neighborhood Peer Effects in Secondary School Enrollment Decisions." *The Review of Economics and Statistics* 91 (4): 695–716.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Nganga, 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.

CBSE. 2022. *Experiential Learning.* Manual. New Delhi, India: Central Board of Secondary Education.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126 (4): 1593–1660.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Nompumelelo Mohohlwane, Stephen Taylor, and Tsegofatso Thulare. 2022. "Can virtual replace in-person coaching? Experimental evidence on teacher professional development and student learning." *Journal of Development Economics* 155 (March): 102815.

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.

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. 2022. *Experimental Evidence on Alternative Policies to Increase Learning at Scale.* Technical report w27298. Cambridge, MA: National Bureau of Economic Research, May.

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

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review,* 1241–1278.

Evans, David K., and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31, no. 2 (August): 242–270.

Evans, David K., and Fei Yuan. 2022. "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis* (April): 01623737221079646.

Fuchs, Lynn S, Nicole Bucka, Ben Clarke, Barbara Dougherty, Nancy C Jordan, Karen S Karp, John Woodward, et al. 2021. *Assisting Students Struggling with Mathematics: Intervention in the Elementary Grades.* Report WWC 2021006. Washington, D.C.: What Works Clearinghouse, March.

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.

Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on NonTest Score Outcomes." *Journal of Political Economy* 126, no. 5 (October): 2072–2107.

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.

Li, Tao, Li Han, Linxiu Zhang, and Scott Rozelle. 2014. "Encouraging classroom peer interactions: Evidence from Chinese migrant schools." *Journal of Public Economics* 111 (March): 29–45.

Loyalka, Prashant, Anna Popova, Guirong Li, and Zhaolei Shi. 2019. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program." *American Economic Journal: Applied Economics* 11, no. 3 (July): 128–154.

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

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *Quarterly Journal of Economics* 134, no. 3 (August): 1627–1673.

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 Education. 2021. *National Initiative for Proficiency in Reading with Understanding and Numeracy (NIPUN BHARAT).* Implementation Guidelines. New Delhi, India: Department of School Education & Literacy, Ministry of Education, Government of India.

Mizala, A., and M. Urquiola. 2013. "School markets: The impact of information approximating schools' effectiveness." *Journal of Development Economics* 103:313–335.

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

Nyqvist, Martina Björkman, and Andrea Guariso. 2021. "Supporting Learning In and Out of School: Experimental Evidence from India." Stockholm, June.

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, David K Evans, Mary E Breeding, and Violeta Arancibia. 2022. "Teacher Professional Development around the World: The Gap between Evidence and Practice." *The World Bank Research Observer* 37, no. 1 (February): 107–136.

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.

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

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.

Schaffner, Julie, Paul Glewwe, and Uttam Sharma. 2021. *Why Programs Fail: Lessons for Improving Public Service Quality from a Mixed-Methods Evaluation of an Unsuccessful Teacher Training Program in Nepal.* Working Paper P21-3. Minneapolis, MN: University of Minnesota.

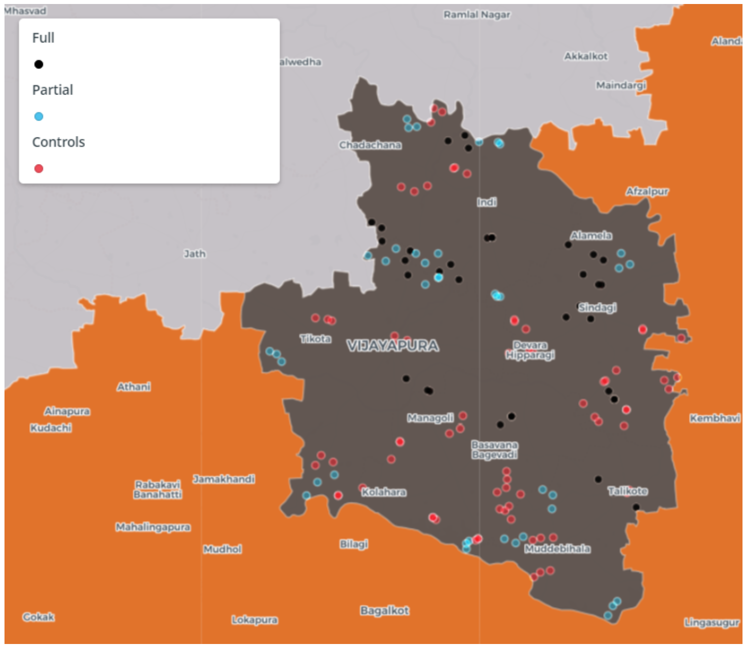
Singh, Abhijeet. 2015. "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages." *Journal of Development Economics* 113:16–32.

The World Bank. 2020. *Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low and Middle Income Countries?* Technical report. Washington, D.C.: The World Bank, October.

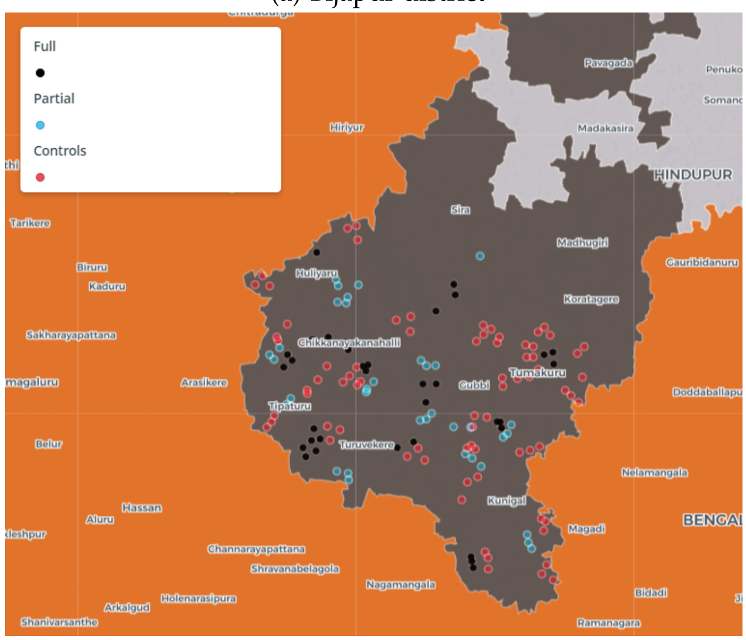
UNESCO Institute for Statistics. 2018. *Data for the Sustainable Development Goals.*

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).

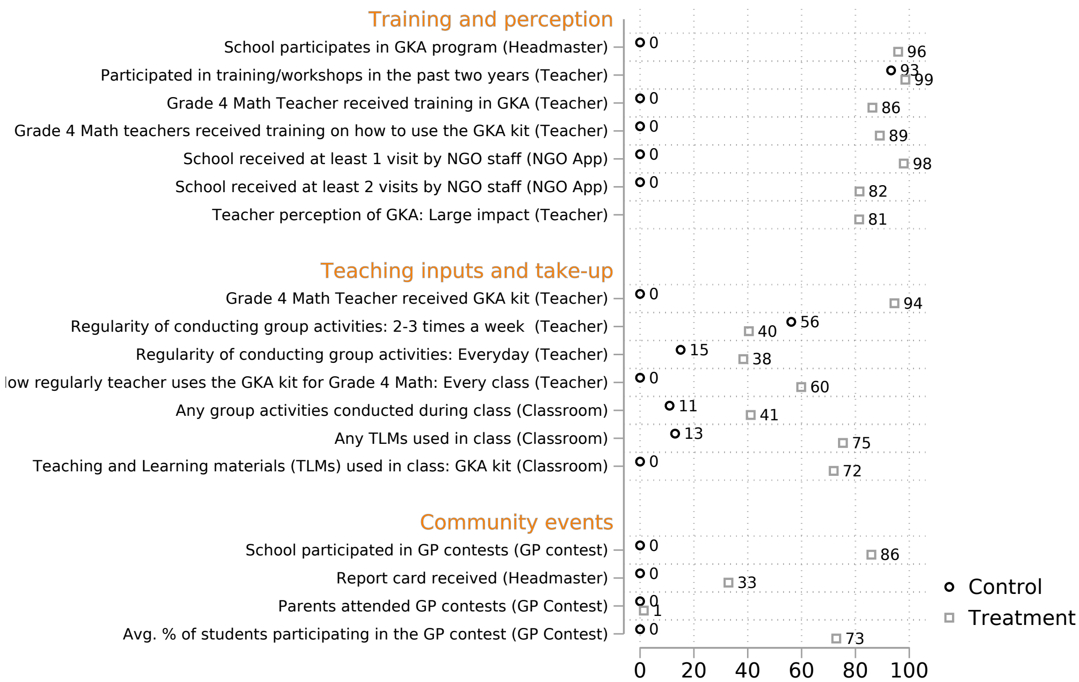


FIGURE II
Implementation fidelity and program take-up

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.

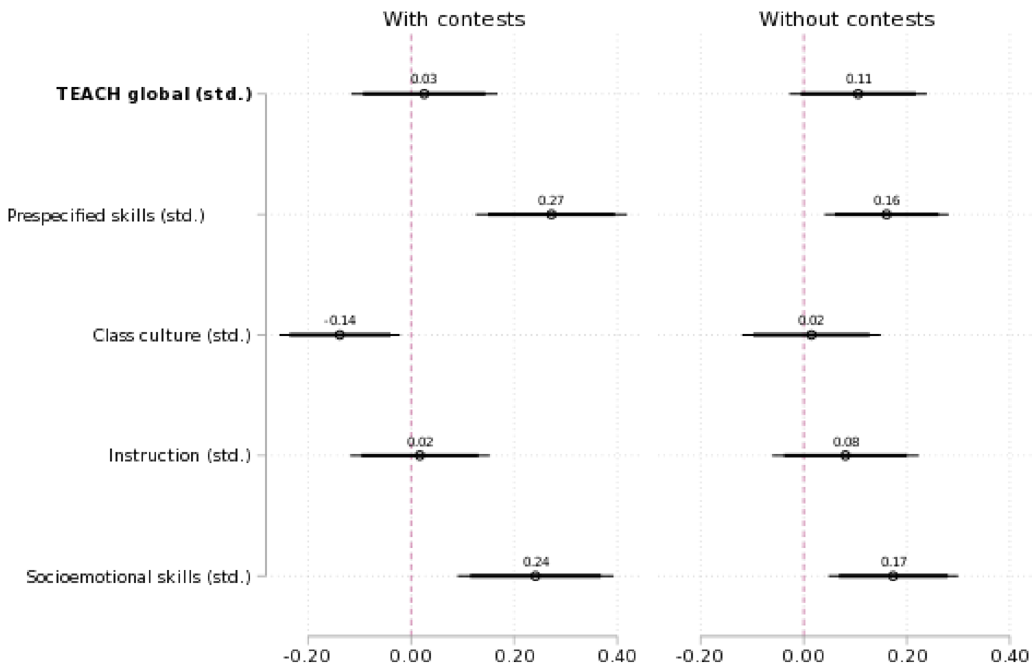
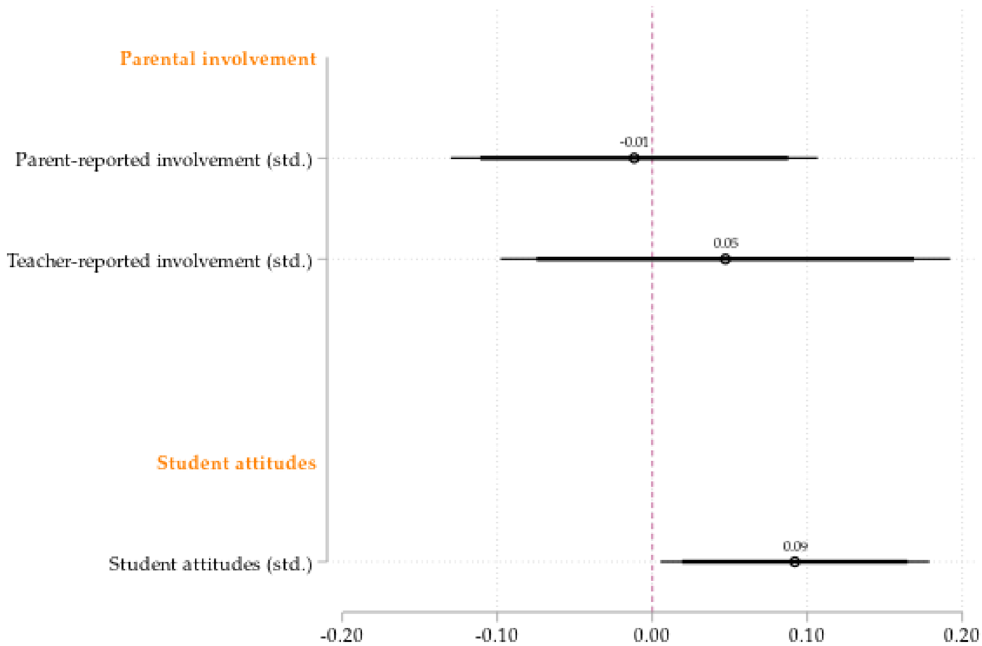
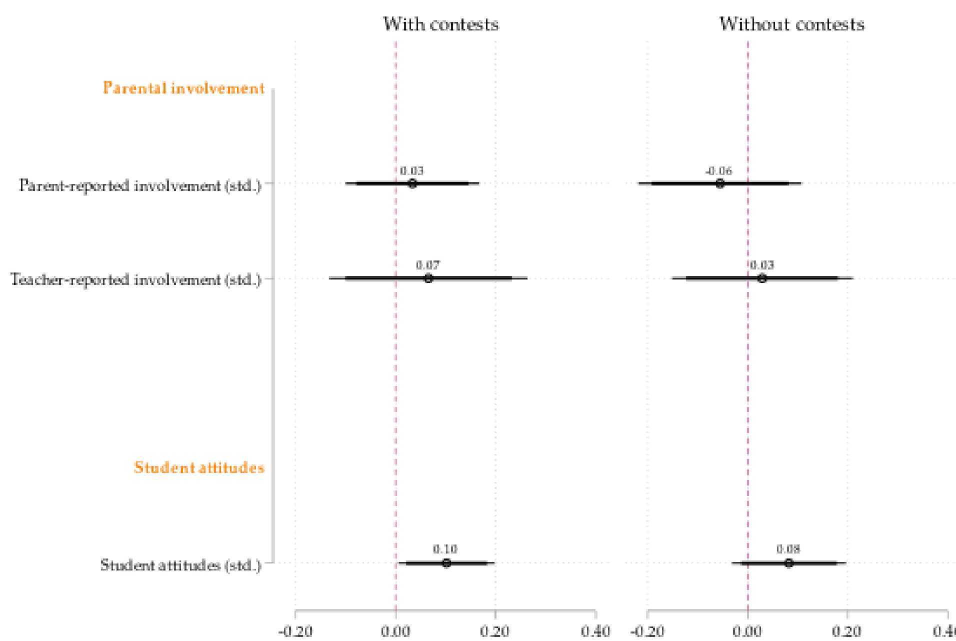


FIGURE III
ITT effects on instructional quality

This figure depicts the program's ITT effects on instructional quality, by program variant, across the entire study period. 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 A4). Table VI presents the effects of adding the contests to the intervention. 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 consist of 1,937 parent interviews, 871 teacher interviews, and 5,649 student interviews, respectively. Thick/thin horizontal bars show 90-/95-percent confidence intervals.

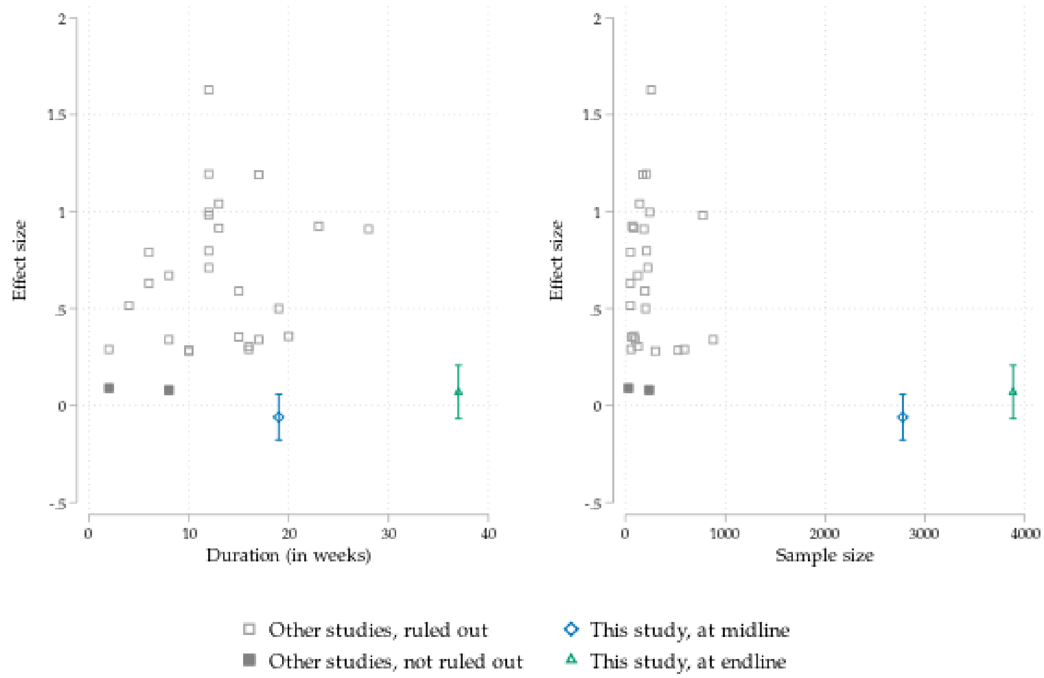


FIGURE V
Comparison with other studies

This figure compares the study's treatment effects with those from other elementary-grade studies of the same pedagogical approach, as identified in a recent systematic review by the What Works Clearinghouse (Fuchs et al. 2021). Each dot represents one study; we average effect sizes if a study reports on effects for multiple outcomes (e.g., for whole numbers computation and whole numbers magnitude understanding). Vertical bars show 95-percent confidence intervals; "ruled out" refers to effect sizes that are greater than the larger upper bound. Panels to the left show the exposure time on the x-axis (in weeks); panels to the right show the study's sample size on the x-axis.

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>=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>=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>=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>=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 (prop.)	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 (prop.)	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 (prop.)	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 Ops. (prop.)	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	0.28	0.29	0.29	0.00	0.00	0.00
				[0.45]	[0.45]	[0.45]	(0.02)	(0.02)	(0.03)
Attrition at endline	1862	1002	1017	0.19	0.23	0.19	0.03*	-0.01	0.04**
				[0.40]	[0.42]	[0.39]	(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; “2pl, std.” refers to the two-parameter logistic item response theory (IRT) model, standardized with respect to the control group at baseline; “prop.” refers to the proportion of test questions answered correctly; “HOTS” and “LOTS” refer to higher- and lower-order thinking skills, respectively. Standard deviations in brackets; standard errors in parentheses (clustered at the Panchayat level). All estimations include randomization strata fixed effects (F.E.s).

TABLE II

ITT EFFECTS ON STUDENT LEARNING

		Control group		ITT effects	
	Baseline mean	Gain to midline	Gain to endline	At midline	At endline
	(1)	(2)	(3)	(4)	(5)
Panel A: Effects on main outcome					
Written test	0.04	0.13*	0.40***	-0.06	0.07
	[0.97]	(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
	[0.30]	(0.01)	(0.01)	(0.01)	(0.01)
ASER>=Subtraction	0.34	0.09***	0.24***	-0.01	0.01
	[0.47]	(0.02)	(0.01)	(0.02)	(0.03)
ASER>=Division	0.09	0.01	0.11***	0.01	0.02
	[0.29]	(0.01)	(0.02)	(0.01)	(0.02)
Panel C: Effects by cognitive domain					
Higher-order	0.04	0.08	0.27***	-0.05	0.03
	[0.98]	(0.08)	(0.08)	(0.06)	(0.07)
Lower-order	0.04	0.10	0.31***	-0.06	0.10
	[0.98]	(0.07)	(0.07)	(0.06)	(0.06)
Panel D: Effects by content domain					
Data	0.38	0.10***	0.11***	-0.02	0.01
	[0.21]	(0.02)	(0.02)	(0.02)	(0.02)
Geometry	0.49	-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 Operations	0.52	-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]). In column 1, the sample consists of students with a written test score at endline. 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.

TABLE III

ITT EFFECTS ON STUDENT LEARNING, BY PROGRAM VARIANT

	Control group			ITT effects at midline			ITT effects at endline		
	Baseline mean	Gain to midline	Gain to endline	Contests vs Control	Materials vs Control	Contests vs Materials	Contests vs Control	Materials vs Control	Contests vs Materials
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Effects on main outcome									
Main outcome	0.04	0.13*	0.40***	-0.10	-0.02	-0.09	0.01	0.12	-0.10
	[0.97]	(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*
	[0.30]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ASER (Baseline)>=Subtraction	0.34	0.09***	0.24***	-0.02	-0.00	-0.02	-0.00	0.02	-0.02
	[0.47]	(0.02)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
ASER (Baseline)>=Division	0.09	0.01	0.11***	0.03	-0.02	0.04**	0.01	0.02	-0.01
	[0.29]	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)
Panel C: Effects by cognitive domain									
Higher-order	0.04	0.08	0.27***	-0.10	-0.01	-0.09	-0.04	0.08	-0.12
	[0.98]	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.09)	(0.07)	(0.11)
Lower-order	0.04	0.10	0.31***	-0.11	-0.01	-0.10	0.06	0.14**	-0.08
	[0.98]	(0.07)	(0.07)	(0.07)	(0.07)	(0.09)	(0.09)	(0.07)	(0.10)
Panel D: Effects by content domain									
Data	0.38	0.10***	0.11***	-0.02	-0.01	-0.01	-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	0.49	-0.00	-0.14***	-0.02	0.01	-0.04*	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 Operations	0.52	-0.03	-0.04**	-0.03	-0.00	-0.03	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]). In column 1, the sample consists of students with a written test score at endline. 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.

TABLE IV
Heterogeneity in ITT effects on student learning

	Control group			ITT effects	
	Baseline mean	Gain to midline	Gain to endline	At midline	At endline
	(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	0.19**	0.51***	-0.09	-0.02
	[0.96]	(0.09)	(0.08)	(0.07)	(0.07)
Male vs Female	-0.13*	0.10*	0.20***	-0.06	-0.16**
	(0.07)	(0.05)	(0.06)	(0.05)	(0.07)
Panel B: By baseline learning level					
Bottom tercile	-1.06	0.61***	0.96***	-0.12	0.07
	[0.55]	(0.05)	(0.04)	(0.08)	(0.08)
Middle tercile	0.01	0.20***	0.41***	-0.05	0.04
	[0.25]	(0.04)	(0.04)	(0.07)	(0.08)
Top tercile	1.05	-0.32***	-0.10**	-0.02	0.09
	[0.49]	(0.04)	(0.04)	(0.08)	(0.09)
Top vs bottom tercile	2.07	-0.94***	-1.06***	0.10	0.03
	(1.48)	(0.09)	(0.11)	(0.11)	(0.10)
Panel C: By district					
Bijapur	-0.04	0.10**	0.30***	-0.13*	0.08
	[0.76]	(0.05)	(0.04)	(0.08)	(0.09)
Tumkur	0.19	0.18***	0.58***	0.04	0.06
	[0.04]	(0.06)	(0.05)	(0.09)	(0.10)
Tumkur vs Bijapur	0.22*	0.08	0.27**	0.18	-0.02
	(0.13)	(0.14)	(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. In column 1, the sample consists of students with a written test score at endline. 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.

TABLE V

Heterogeneity in ITT effects on student learning, by program variant

	Control group			ITT effects at midline			ITT effects at endline		
	Baseline mean	Gain to midline	Gain to endline	Contests vs Control	Materials vs Control	Contests vs Materials	Contests vs Control	Materials vs Control	Contests vs Materials
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: By gender									
Female	0.12	0.09	0.31***	-0.08	0.01	-0.09	0.09	0.18**	-0.09
	[0.98]	(0.07)	(0.08)	(0.08)	(0.07)	(0.09)	(0.10)	(0.09)	(0.12)
Male	-0.05	0.19**	0.51***	-0.14*	-0.06	-0.08	-0.08	0.04	-0.12
	[0.96]	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.08)	(0.10)
Male vs Female	-0.13*	0.10*	0.20***	-0.06	-0.07	0.01	-0.17**	-0.15*	-0.02
	(0.07)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)
Panel B: By baseline learning level									
Bottom tercile	-1.06	0.61***	0.96***	-0.13	-0.10	-0.03	0.03	0.12	-0.09
	[0.55]	(0.05)	(0.04)	(0.10)	(0.10)	(0.11)	(0.11)	(0.10)	(0.13)
Middle tercile	0.01	0.20***	0.41***	-0.07	-0.03	-0.03	0.02	0.06	-0.04
	[0.25]	(0.04)	(0.04)	(0.08)	(0.09)	(0.10)	(0.11)	(0.08)	(0.11)
Top tercile	1.05	-0.32***	-0.10**	-0.11	0.07	-0.18	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	-0.94***	-1.06***	0.02	0.17	-0.15	-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	0.10**	0.30***	-0.15	-0.11	-0.04	0.00	0.16	-0.16
	[0.76]	(0.05)	(0.04)	(0.10)	(0.09)	(0.11)	(0.12)	(0.11)	(0.15)
Tumkur	0.19	0.18***	0.58***	-0.03	0.10	-0.13	0.04	0.07	-0.03
	[0.04]	(0.06)	(0.05)	(0.09)	(0.12)	(0.12)	(0.12)	(0.12)	(0.13)
Tumkur vs Bijapur	0.22*	0.08	0.27**	0.12	0.21	-0.09	0.04	-0.09	0.13
	(0.13)	(0.14)	(0.14)	(0.13)	(0.15)	(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. In column 1, the sample consists of students with a written test score at endline. 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.

TABLE VI

ITT EFFECTS OF ADDING COMMUNITY CONTESTS TO THE INTERVENTION, ON INSTRUCTIONAL QUALITY

	Round 2		Round 3		Round 4		Rounds 3-4 (pooled)	
	(1) All rounds	(2) Round-1 controls	(3) All rounds	(4) Round-1 controls	(5) All rounds	(6) Round-1 controls	(7) All rounds	(8) Round-1 controls
TEACH global (std.)	0.14	0.07	-0.30**	-0.38***	-0.26**	-0.30**	-0.28***	-0.34***
	(0.13)	(0.14)	(0.12)	(0.13)	(0.13)	(0.12)	(0.09)	(0.08)
Pre-specified skills (std.)	0.25**	0.26**	-0.15	-0.16	-0.02	0.00	-0.09	-0.08
	(0.12)	(0.13)	(0.14)	(0.14)	(0.14)	(0.14)	(0.11)	(0.10)
Class culture (std.)	0.18	0.10	-0.48***	-0.59***	-0.30*	-0.38**	-0.40***	-0.49***
	(0.15)	(0.17)	(0.14)	(0.14)	(0.16)	(0.16)	(0.09)	(0.09)
Instruction (std.)	-0.05	-0.11	-0.15	-0.21*	-0.14	-0.15	-0.15	-0.18*
	(0.13)	(0.13)	(0.12)	(0.12)	(0.17)	(0.15)	(0.10)	(0.09)
Socioemotional skills (std.)	0.19	0.20*	-0.02	-0.03	-0.07	-0.05	-0.04	-0.04
	(0.12)	(0.12)	(0.15)	(0.14)	(0.15)	(0.15)	(0.11)	(0.10)

Notes: This table presents the ITT effects of adding contests to the program, on instructional quality. The first row reflects effects on the overall *Teach* index; the following three rows reflect effects on its three subdimensions. 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 A4). The estimation sample consists of 1,615 classroom observation ratings. Odd models ("All obs.") include all ratings and estimate round-specific effects with interaction terms; even models ("Round-1 controls") drop Round-1 observations and use their ratings as school-level controls.

TABLE VII

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	0.07	0.07	0.07	0.06	0.09	0.07	0.07
	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	[0.45]
Effect with events	0.01	0.00	0.01	0.02	0.02	0.02	0.03	0.01	0.01
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	[0.91]
Effect without events	0.12*	0.08	0.14**	0.12	0.12	0.11	0.13*	0.12*	0.12
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	[0.28]
Effect without events, among girls	0.20**	0.14*	0.20**	0.18**	0.18**	0.17*	0.20**	0.18**	0.18
	(0.09)	(0.08)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	[0.14]
Effect without events, among boys	0.03	0.02	0.06	0.04	0.04	0.02	0.05	0.05	0.04
	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	[0.74]

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.

TABLE VIII

School value-added on student learning and instructional quality

	(1)	(2)	(3)	(4)	(5)	(6)
TEACH global		-0.03				
		(0.03)				
Pre-specified skills			-0.05**			
			(0.02)			
Class culture				-0.02		
				(0.03)		
Instruction					-0.02	
					(0.03)	
Socioemotional skills						-0.02
						(0.02)
Teacher age	0.00	-0.00	-0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher is female	0.01	0.00	0.00	0.01	0.01	0.01
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Teacher years at school	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher holds teaching degree	0.04	0.04	0.04	0.04	0.04	0.04
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
R^2	0.00	0.01	0.02	0.01	0.01	0.01

Notes. This table shows results from regressions of school value-added on teacher characteristics and measures of instructional quality. The dependent variable is the school's fixed effect's deviation from the Panchayat mean, from a regression of student endline scores on school fixed effects, baseline scores, and a vector of control variables selected via Lasso ($n = 292$). Regressions control for the treatment indicators (results not shown). Standard errors in parentheses (clustered at the Panchayat level).

Appendix

A Additional Figures and Tables

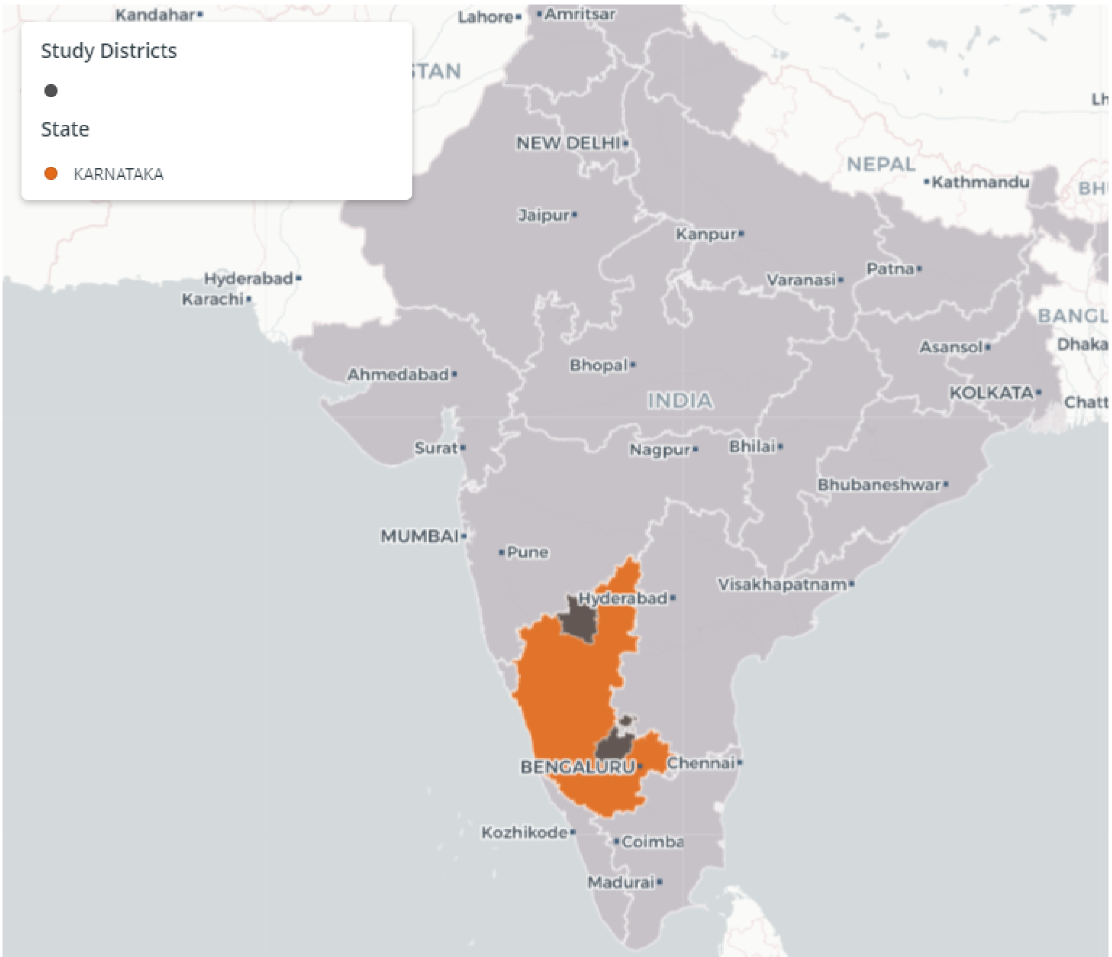


Figure A1
Location of the study

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

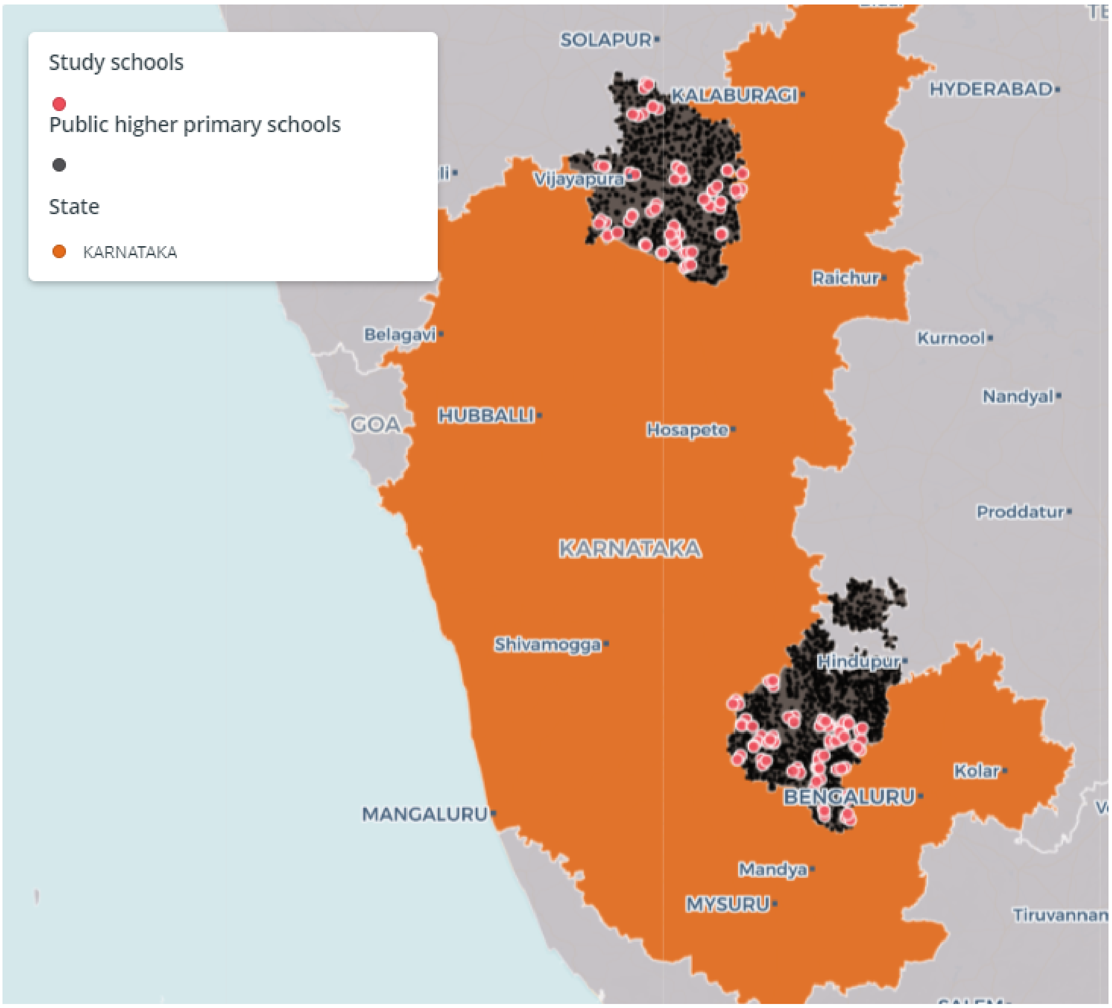


Figure A2
Random sampling of GPs and primary schools in study districts

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.

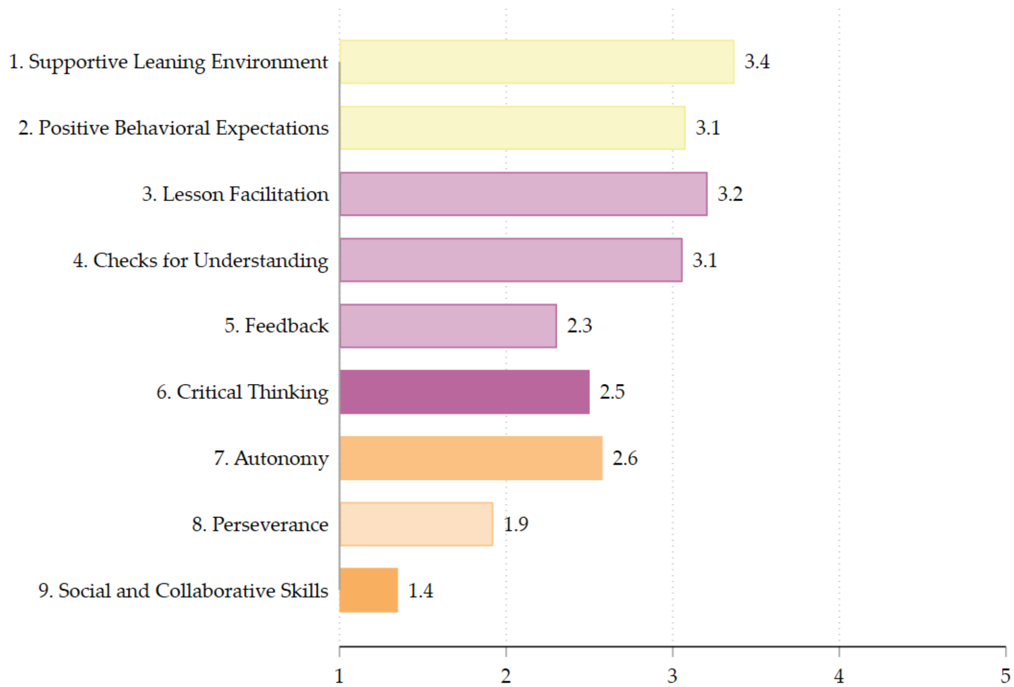


FIGURE A3
Description of teaching quality in the control group

This figure depicts mean *Teach* classroom observation scores, for the control group, as measured during process monitoring rounds. Sub-domains one and two relate to classroom culture, sub-domains three to six relate to instruction, and sub-domains seven to nine relate to the promotion of socio-emotional skills. Solid bars indicate the three pre-specified sub-domains that were jointly expected to relate to the intervention: critical thinking, autonomy, and social and collaborative skills. Ratings range from one ("low") to five ("high").

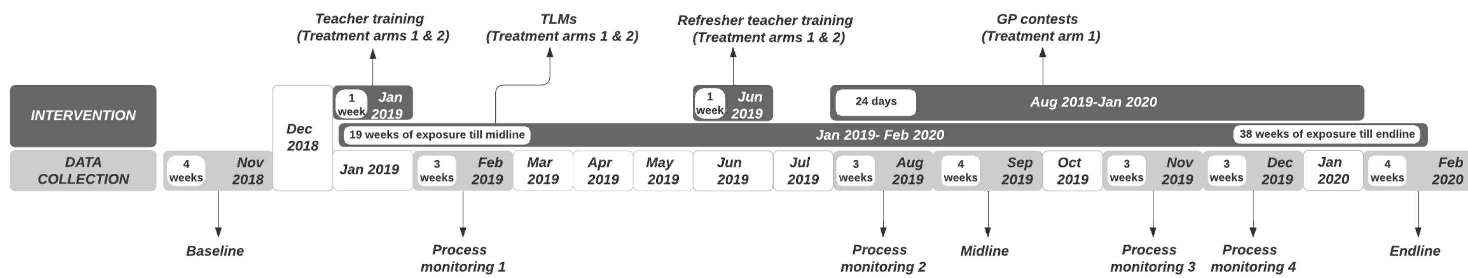
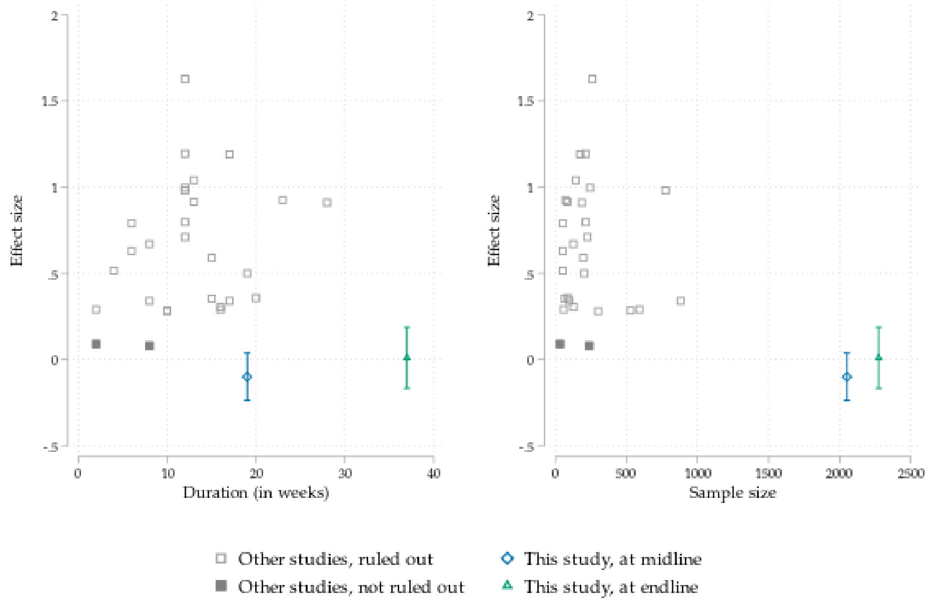
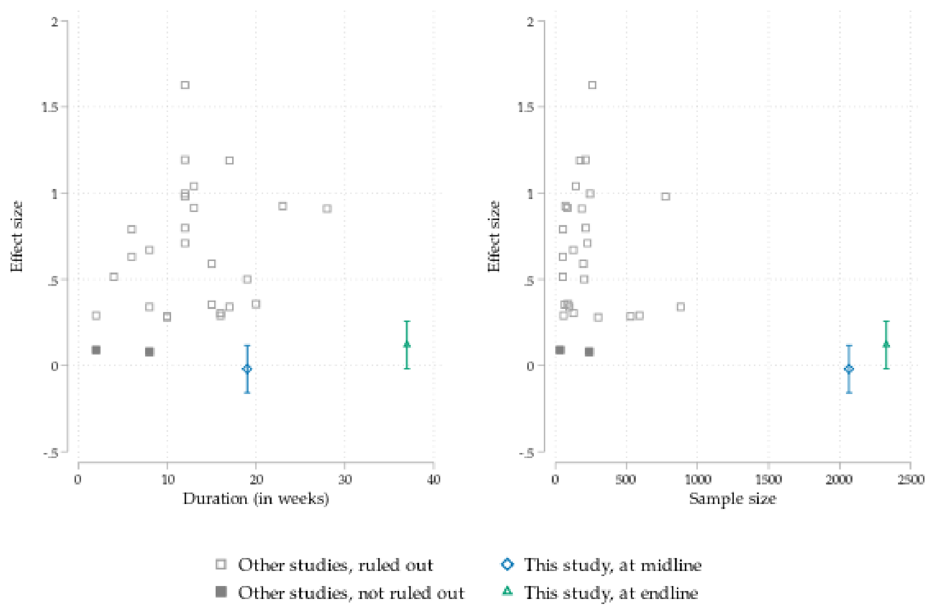


FIGURE A4
Study timeline

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.



(a) Without contests



(b) With contests

FIGURE A5
Comparison with other studies, by program variant

This figure compares the study's treatment effects with those from other elementary-grade studies of the same pedagogical approach, as identified in a recent systematic review by the What Works Clearinghouse (Fuchs et al. 2021). Each dot represents one study; we average effect sizes if a study reports on effects for multiple outcomes (e.g., for whole numbers computation and whole numbers magnitude understanding). Subfigure (a) reports on effect sizes for the program variant without contests; Subfigure (b) reports on effect sizes for the program variant with contests. Vertical bars show 95-percent confidence intervals; "ruled out" refers to effect sizes that are greater than the larger upper bound. Panels to the left show the exposure time on the x-axis (in weeks); panels to the right show the study's sample size on the x-axis.

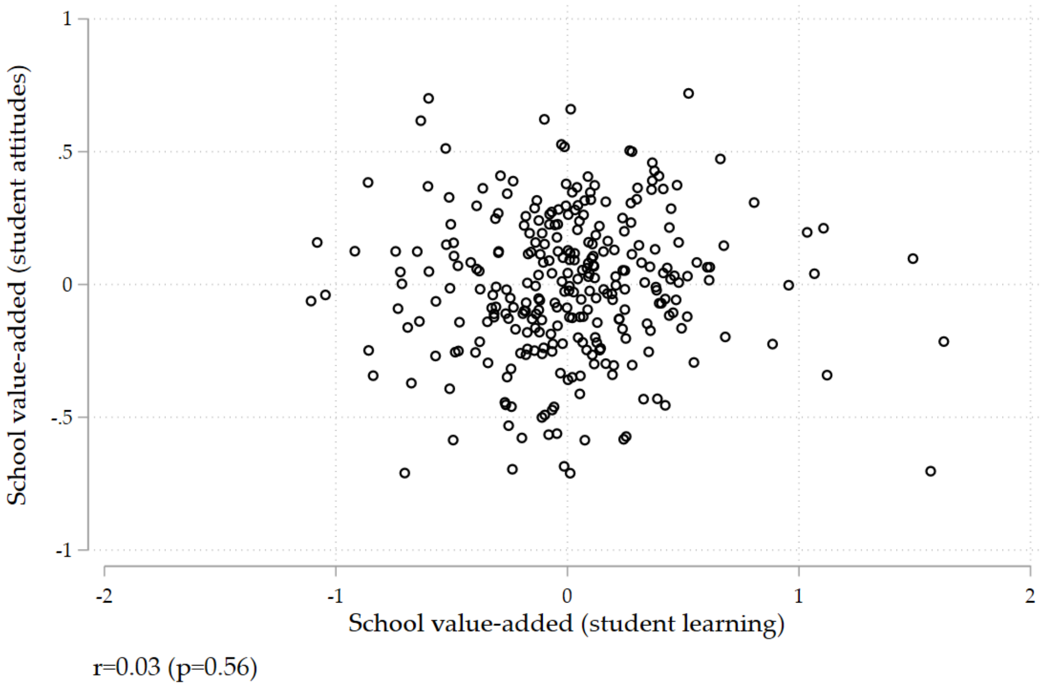


Figure A6
Correlation between school value-added on learning vs on student attitudes

This figure provides a scatter plot, and the correlation, of schools' value-added on student learning and student attitudes, respectively. Each dot is a school's fixed effect's deviation from the Panchayat mean, from a regression of student endline scores or attitudes towards mathematics on school fixed effects, baseline scores, and a vector of control variables selected via Lasso ($n = 292$).

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)	1492	772	821	9.13	9.13	9.14	-0.01	0.02	-0.03
				[0.54]	[0.54]	[0.58]	(0.03)	(0.03)	(0.03)
Female	1500	774	828	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.)	1500	774	828	0.04	-0.03	-0.03	-0.03	-0.08	0.05
				[0.97]	[0.98]	[0.96]	(0.05)	(0.08)	(0.07)
ASER >=1-digit	1500	774	828	0.99	0.98	0.99	-0.01	-0.01	-0.00
				[0.09]	[0.12]	[0.11]	(0.01)	(0.01)	(0.01)
ASER >=2-digit	1500	774	828	0.90	0.88	0.92	-0.02	0.02	-0.03**
				[0.30]	[0.33]	[0.27]	(0.02)	(0.01)	(0.01)
ASER >=Subtraction	1500	774	828	0.34	0.33	0.34	-0.01	-0.01	0.01
				[0.47]	[0.47]	[0.47]	(0.02)	(0.02)	(0.02)
ASER >=Division	1500	774	828	0.09	0.10	0.10	0.01	0.01	0.00
				[0.29]	[0.30]	[0.30]	(0.01)	(0.01)	(0.02)
Math, HOTS (2pl, std.)	1500	774	828	0.04	-0.04	-0.04	-0.03	-0.08	0.05
				[0.98]	[1.00]	[0.98]	(0.06)	(0.09)	(0.08)
Math, LOTS (2pl, std.)	1500	774	828	0.04	-0.02	-0.04	-0.03	-0.09	0.06
				[0.98]	[0.96]	[0.96]	(0.05)	(0.07)	(0.07)
Data (prop.)	1500	774	828	0.38	0.35	0.35	-0.02	-0.02	0.00
				[0.21]	[0.21]	[0.21]	(0.01)	(0.02)	(0.02)
Geometry (prop.)	1500	774	828	0.49	0.48	0.48	-0.00	-0.02	0.02
				[0.29]	[0.28]	[0.29]	(0.02)	(0.03)	(0.02)
Number Sense (prop.)	1500	774	828	0.61	0.59	0.59	-0.01	-0.01	-0.00
				[0.27]	[0.28]	[0.27]	(0.02)	(0.02)	(0.02)
Whole Number Ops. (prop.)	1500	774	828	0.52	0.51	0.50	-0.00	-0.03	0.02
				[0.29]	[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; "2pl, std." refers to the two-parameter logistic item response theory (IRT) model, standardized with respect to the control group at baseline; "prop." refers to the proportion of test questions answered correctly; "HOTS" and "LOTS" refer to higher- and lower-order thinking skills, respectively. 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).

TABLE A2

ITT effects of adding community contests to the intervention, on instructional quality (by indicator)

	Round 2		Round 3		Round 4		Rounds 3-4 (pooled)	
	(1) All rounds	(2) Round-1 controls	(3) All rounds	(4) Round-1 controls	(5) All rounds	(6) Round-1 controls	(7) All rounds	(8) Round-1 controls
Learning environment	0.11	0.02	-0.22	-0.30**	-0.18	-0.22	-0.20**	-0.26***
	(0.15)	(0.17)	(0.14)	(0.14)	(0.15)	(0.15)	(0.10)	(0.09)
Behavioral expectations	0.19	0.16	-0.56***	-0.64***	-0.29**	-0.35***	-0.43***	-0.50***
	(0.12)	(0.13)	(0.16)	(0.16)	(0.14)	(0.13)	(0.09)	(0.09)
Lesson facilitation	0.08	0.07	-0.28**	-0.28**	0.20	0.22	-0.05	-0.03
	(0.12)	(0.12)	(0.12)	(0.12)	(0.16)	(0.16)	(0.10)	(0.10)
Checks for understanding	-0.15	-0.21	0.11	0.08	-0.28*	-0.29*	-0.08	-0.10
	(0.14)	(0.13)	(0.13)	(0.13)	(0.17)	(0.16)	(0.10)	(0.08)
Feedback	-0.05	-0.12	-0.06	-0.13	-0.18	-0.21	-0.12	-0.17
	(0.17)	(0.17)	(0.18)	(0.18)	(0.17)	(0.15)	(0.12)	(0.12)
Critical thinking	-0.00	0.01	-0.10	-0.12	-0.02	-0.02	-0.06	-0.07
	(0.11)	(0.12)	(0.12)	(0.12)	(0.14)	(0.14)	(0.10)	(0.10)
Autonomy	0.24*	0.25**	-0.21	-0.23	-0.08	-0.04	-0.15	-0.14
	(0.12)	(0.12)	(0.16)	(0.16)	(0.14)	(0.15)	(0.10)	(0.10)
Perseverance	-0.12	-0.11	0.05	0.04	-0.11	-0.12	-0.02	-0.04
	(0.11)	(0.10)	(0.17)	(0.15)	(0.14)	(0.14)	(0.12)	(0.11)
Social and collaborative skills	0.24	0.30	0.13	0.18	0.09	0.13	0.11	0.15
	(0.19)	(0.18)	(0.15)	(0.14)	(0.27)	(0.26)	(0.16)	(0.14)

Notes: This table presents the ITT effects of adding contests to the program, on each of the individual indicators of instructional quality. 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 A4). The estimation sample consists of 1,615 classroom observation ratings. Odd models ("All obs.") include all ratings and estimate round-specific effects with interaction terms; even models ("Round-1 controls") drop Round-1 observations and use their ratings as school-level controls.

TABLE A3

School value-added on student attitudes and instructional quality

	(1)	(2)	(3)	(4)	(5)	(6)
TEACH global		0.02				
		(0.02)				
Pre-specified skills			0.00			
			(0.02)			
Class culture				0.04**		
				(0.02)		
Instruction					0.02	
					(0.02)	
Socioemotional skills						0.01
						(0.01)
Teacher age	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher is female	-0.01	-0.01	-0.01	-0.00	-0.01	-0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Teacher years at school	0.01**	0.01**	0.01**	0.01**	0.01*	0.01**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher holds teaching degree	-0.01	-0.00	-0.01	0.01	-0.00	-0.01
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
R^2	0.02	0.02	0.02	0.04	0.02	0.02

Notes. This table shows results from regressions of school value-added on teacher characteristics and measures of instructional quality. The dependent variable is the school's fixed effect's deviation from the Panchayat mean, from a regression of student attitudes towards mathematics on school fixed effects, baseline scores, and a vector of control variables selected via Lasso ($n = 292$). Regressions control for the treatment indicators (results not shown). Standard errors in parentheses (clustered at the Panchayat level).

TABLE A4

Baseline test scores and instructional quality

	(1)	(2)	(3)	(4)	(5)	(6)
TEACH global		0.10***				
		(0.03)				
Pre-specified skills			0.08***			
			(0.03)			
Class culture				0.03		
				(0.03)		
Instruction					0.12***	
					(0.04)	
Socioemotional skills						0.03
						(0.03)
Teacher age	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher is female	-0.04	-0.02	-0.03	-0.03	-0.04	-0.03
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Teacher years at school	0.00	0.00	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Teacher holds teaching degree	0.41***	0.43***	0.41***	0.42***	0.42***	0.41***
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
R^2	0.04	0.07	0.06	0.04	0.08	0.04

Notes. This table shows results from regressions of baseline test scores and measures of instructional quality. The dependent variable is the school's average baseline test score in math ($n = 292$). Regressions control for the treatment indicators (results not shown). Standard errors in parentheses (clustered at the Panchayat level).

B The Intervention

The two main components of the intervention consist of (1) teaching inputs for activity-based instruction, and related teacher training, and (2) community contests. The following describes these two components in greater detail. Here, we focus on the program's design; in the main text, we provide detailed information on observed implementation fidelity and program take-up. Appendix Figure B1 summarizes the program's Theory of Change.

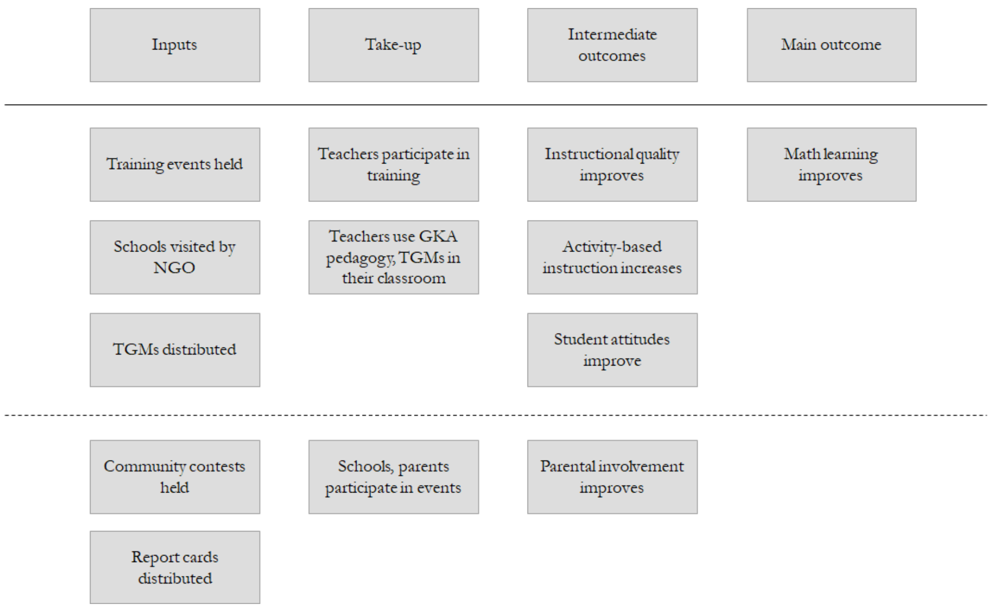


Figure B1
Theory of Change

This figure summarizes the program's Theory of Change. Items below the dashed line refer to program components that are provided only by the program version with community events. GKA refers to the Ganitha Kalika Andolana program. TGM refers to Teaching and Learning Materials.

Teaching inputs for activity-based instruction, and related training

The pedagogical approach of the GKA perceives of learning as an active and social process that happens when a child encounters hands-on, activity-based experiences. Group activities and collaborative peer learning is a major part of this approach. Teachers are expected to integrate this pedagogy in their usual mathematics classes. To this end, they receive additional teaching inputs, as well as training.

The additional teaching inputs consist of a box of 19 Teaching Learning Materials (TLMs) that can be used to teach 20 mathematical concepts (the "GKA kit"). These TLMs are sensory in nature and enable the children to learn these concepts through a range of visual and tactile experiences. The kit consists of a series of items such as the abacus, a series of shapes and letters, clocks, and measuring kits. The GKA kit also contains the training manual that ties the TLMs to the various mathematical concepts teachers are expected to teach (as per the official state curriculum). The manual includes "math concept cards" to help the teacher ask conceptual questions that encourage the child to use their problem-solving skills. Finally, the GKA kit also provides guidance on how to facilitate group learning activities in the classroom setting. Appendix Table B1 lists the contents of teaching and learning materials along with mathematical concepts students are expected to learn, as per the state curriculum.

In order to use the TLMs and the teaching manual effectively, each teacher is trained both off-site and on-site. During the study period, teachers were invited to participate in two five-day off-site trainings (an initial training and a refresher training). On-site training mainly consists of follow-up monitoring visits to schools, through Akshara staff. During the study period, one staff member was allotted to approximately 75 schools (as "field coordinator").

Community contests

The community contests are mathematics contests that are held at the Gram Panchayat level ("GP contests"). The objective of these contests is to build awareness among the community about children's learning levels. The program's theory of change posits that, if the community becomes aware of children's low learning levels, community members will demand higher educational quality and improved child learning.

The contests involve testing the children in mathematics concepts; the assessment itself takes place in a public setting. Once the assessment is conducted, the top three scorers in grades 4, 5, and 6 are given prizes.[25] At the end of the assessment, discussions with attendees seek to increase public awareness of, and greater demand for, quality education. As a follow up, a report card is also created for distribution to leaders at the village and block levels (providing aggregate statistics on the level of participation and student achievement).

25. In contrast, weak performers are not singled out individually.

TABLE B1

Teaching and learning materials and related mathematical concepts

Teaching and learning materials	**Concepts**
Abacus with rings	Numbers, place value, addition, subtraction, patterns, data handling, factors and multiples
Base-10 block (yellow cubes, blue rods, green plates and red block)	Numbers, place value, addition, subtraction, multiplication, division, patterns, data handling, factors and multiples, area and perimeter, weight
Clock	Time, addition, subtraction, angles, multiples of 5
Decimal place value strips	Place value, decimals
Decimal set	Decimals
Dice	Numbers, place value, addition, subtraction, multiplication, division, place value, 3D shapes, pattern, data handling
Fraction shapes	Fractions, 2D shapes, angles, patterns and symmetry
Fraction strips	Fractions
Geo solids with nets	3D shapes
Geo-board	2D shapes, area and perimeter
Measuring tape	Length, fractions, decimals, area and perimeter
Number line and clothes clips	Numbers, addition, subtraction, multiplication, division, factors and multiples
Place value mat and decimal place value mat	Numbers, place value, addition, subtraction
Place value strips	Place value, decimals
Play money and coins	Numbers, money, addition, subtraction, multiplication, division, place value, decimals
Protractor and angle measure	Angles
Square counters	Numbers, addition, subtraction, multiplication, division, fractions, patterns, data handling, factors and multiples
Tangram	2D shapes, patterns, angles
Weighting balance	Weight, volume, addition, subtraction

Notes. This table lists the teaching and learning materials (TLMs) included in the kit, along with mathematical concepts teachers are expected to teach (as per the state curriculum).
Source. Akshara Foundation teacher handbook.

C 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.[26] Secondly, for each round of assessments, we reviewed the test questions with an external panel of subject matter experts.[27] Third, we mapped each test question to the official schoolbooks used 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.

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

27. The panel consisted of former teachers and curriculum experts. The panel did not include staff of the implementing organization.

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 C1 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 C2 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 C1 displays the a and b parameters for the 32 test questions, as per a two parameter logistic (2pl) item response theory (IRT) model.[28] 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.[29] From this analysis, we conclude that our test scores are informative over a wide range of student ability in this 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 ($C\alpha$) is a widely used measure of reliability in psychometric testing. The $C\alpha$ 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 $C\alpha$ 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 $C\alpha$ 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 infor-

28. A three parameter (3PL) model did not converge for the baseline data.
29. 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").

mation 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 C3 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 C1
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 two-parameter logistic (2pl) item response theory (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.

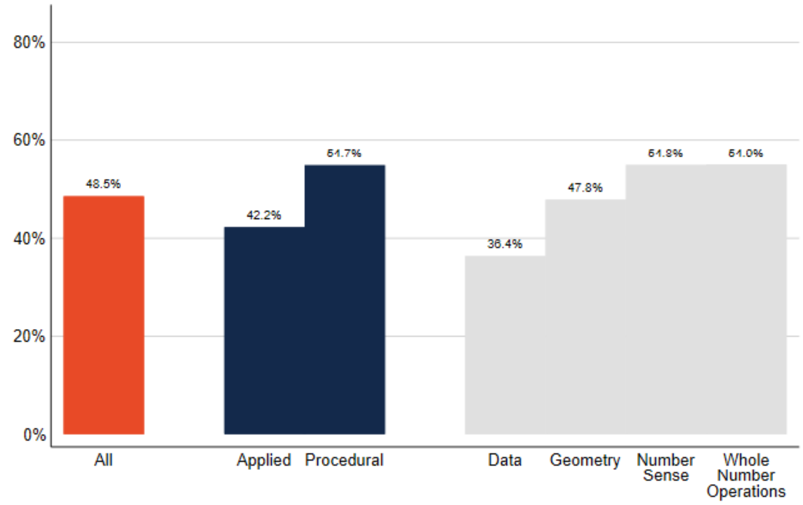


FIGURE C1
Mean percentage of items solved correctly (Baseline)

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

D Statistical Methods

Randomization

We repeated the study's randomization procedure ten times, to select the one with the greatest balance. To do this, we considered the percentage of questions students answered correctly on the oral baseline test and students' learning level as per this test. We then calculated t-statistics for the difference of these two variables across the two groups of GPs. 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. See Bruhn and McKenzie (2009), who refer to this re-randomization approach as the "min-max method."

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 randomization strategy and choosing a conservative number (ten) of re-randomizations. Bruhn and McKenzie (2009) moreover suggest adding the variables used for balancing as covariates in models of program impacts; our models do include students' oral and written baseline perfor-

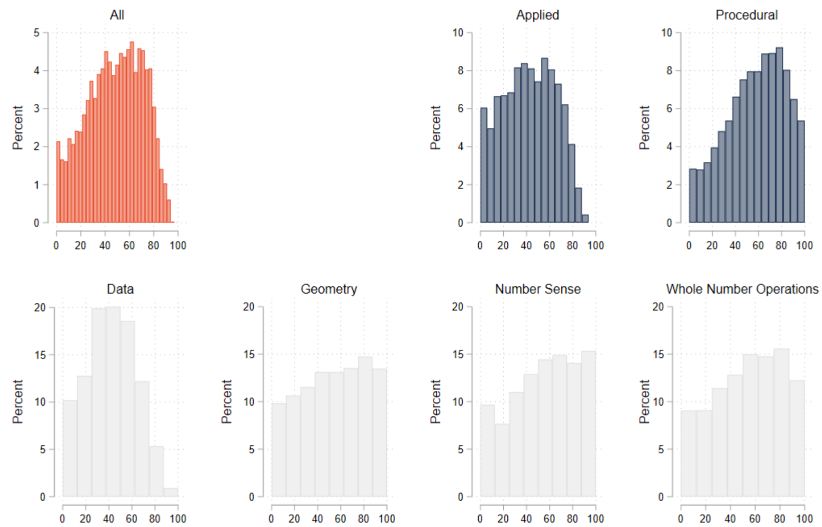


Figure C2
Distribution of percentage of items solved correctly (Baseline)

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

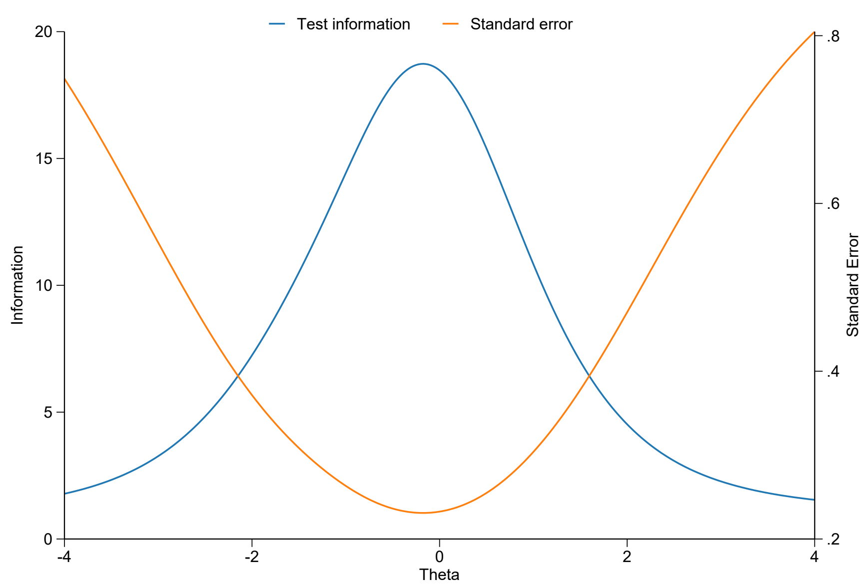


Figure C3
Test information function (TIF)

This figure provides the test information function, and corresponding standard errors of measurement, for the baseline as per a two-parameter logistic (2pl) item response theory (IRT) model.

mance as covariates.

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,[30] and report on program outputs (see Section II.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 this program and the grade levels investigated in our research.

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

30. In the experimental literature, some 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.

(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).[31]

Attrition. We pre-registered three ways to investigate attrition. First, we check whether it is systematically related to treatment status, through tests of differential attrition rates and of selective attrition.[32] Second, we employ two robustness checks: inverse-probability weighting (IPW) and Lee (2009) bounds. 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 two strata that had one school with zero grade-four attendance at baseline.

Multiple outcomes and multiple hypothesis testing

As pre-registered, 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.

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

32. 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).