The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling*

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

June 22, 2023

There is substantial emphasis on improving classroom practices, especially through activity-based instruction, as a way to improve the productivity of schooling. We study a large program that seeks to improve mathematics learning in government primary schools in India. Through a cluster-randomized trial, we find the program increased activity-based instruction but yielded only muted impacts on learning. We provide a potential explanation: School value-added models suggest a negative relationship between activity-based instruction and test score gains. Our findings are robust to adding a community-engagement component to the intervention. These results highlight the limitations of activity-based instruction programs to improve school productivity.

*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.

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 productivity in government primary schools. Government spending on education in India more than doubled between 2006 and 2013 (in constant PPP dollars; 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.

Activity-based instruction is one approach that has recently gained prominence and is starting to be adopted in many developing countries. Activity-based instruction views learning as an active and social process that works best when a child engages in hands-on experiences, often with small groups of other children. Indeed, a recent systematic review by an expert panel of 28 studies from developed countries provides strong support for this pedagogical approach (Fuchs et al. 2021). Moreover, this approach aligns with India's recent National Education Policy 2020, which promotes activity-based, experiential learning in primary grades (Ministry of Education 2021).

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. The program also conducts community-led contests that convene stakeholders to witness the mathematical performance of school children. It is a collaboration between the state government and an Indian non-governmental organization that includes a phased scale-up to all 44,000 government 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 isolate the effect of the pedagogical intervention, we conducted a second randomization and removed the community contests from half of the 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 documenting adherence to treatment assignment, implementation fidelity, and changes in teachers' instructional practices. We find that: 1. All program schools received the additional teaching inputs; 2. Almost all 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 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 three sets of results. First, we show the intention-to-treat (ITT) effects of promoting activity-based instruction on student learning. After 13 months of the program, we estimate an average impact of 0.12 standard deviations (SDs) of the distribution of test scores on students' mathematics skills, although not statistically significant at conventional levels of significance ($p = 0.11$). The program raised girls' math scores by 0.18 SDs ($p < 0.05$), but had no effect on boys (0.04 SDs). We compare our estimates with those of the aforementioned systematic review of other rigorous studies that investigated the same pedagogical approach. We show that, while expert opinion (based on results from developed countries) 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, we explore reasons for these muted impacts of activity-based instruction on student learning. 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 effective-

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

ness in improving student attitudes toward mathematics. We find that the teaching practices promoted by the program are negatively correlated with school effects on test scores, and we find no association with school effects on student attitudes). We also show that these two types of school effects are orthogonal to each other.

Lastly, to explore whether the intervention's (lack of) effectiveness depends on complementarities with another input, we investigate whether adding community contests improves impacts on child learning. Contrary to expectations, 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 0.06 SDs or more are ruled out at 95 percent confidence). In fact, adding community contests to the intervention led to sizeable negative effects on classroom culture and created a less-supportive learning environment for students in the study periods after the community contests were conducted (-0.40 to -0.49 SDs).

Our contribution is threefold. First, there is substantial emphasis on improving classroom practices, especially through activity-based instruction, as a way to improve the productivity of schooling. Our results show the limitations of such activity-based instruction as we rule out the positive effects found by a large body of related efficacy trials from the United States (Fuchs et al. 2021). Our results also diverge from the positive findings from a primary-school intervention that aims to increase students' curiosity in Turkey (Alan and Mumcu 2022) and the promising results of another intervention that promotes "learning to learn" principles and "conceptual learning" in primary schools in Uganda (Ashraf, Banerjee, and Nourani 2021).

Second, as prior evidence on, and the push for, activity-based instruction primarily builds on research from the United States and Europe, our study also speaks to the pitfalls of generalizing across vastly different contexts. Few papers evaluate, at scale, a composite of "best practices" widely recommended for adoption by education practitioners but having little evidence of ever working at scale outside of high-income settings. It is well known that such educational interventions frequently lose their effectiveness at scale if they are not adopted with fidelity (Vivalt 2020). For example, Angrist and Meager (2022) review the effectiveness of a targeted instruction intervention ("Teaching at the Right Level") and document vastly different

impacts depending on the program's delivery model and implementation fidelity. Muralidharan and Singh (2020) evaluate a large-scale educational reform that appeared effective based on administrative measures of compliance but did not affect classroom practices. In contrast, our paper documents a case of high levels of implementation fidelity *and* impacts on the prescribed dimensions of instructional quality yet muted impacts on student learning. Moreover, the baseline levels of school productivity and activity-based instruction are low in India, and so the conditions for generalizability were high in the given study environment (cf. Bates and Glennerster 2017). This suggests that the given intervention failed to generalize to the context of Indian government schools.

Finally, we contribute 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 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 analyses add to evidence that one such practice (activity-based instruction) may reduce student learning (Berlinski and Busso 2017). 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).

1. Background and Intervention

1.1. 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 (Ministry of Home Affairs 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; for example, 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.

1.2. Intervention

We conducted this study cooperating with the Akshara Foundation, a large NGO dedicated to ensuring quality preschool and primary education in India. Founded in 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 the 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 scaling it up to all of the state's 44,000 Government primary schools in a phased manner. In 2017, another Indian state, Odisha, began implementing 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.

1.2.1. 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.[2] 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 primary school teachers. These off-site training sessions are held during the state's scheduled in-service 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 coordinator supports the teachers as they implement this new teaching method.

1.2.2. Community contests

The second component is the community contests. These Gram Panchayat Mathematics Contests ("GP contests") convene stakeholders to observe students' mathematical performance. They are intended to encourage parent engagement and community participation, which can pressure 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.

2. 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.

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 NGO initiates these contests in participating GPs, the GP and other local sources pay for all operational expenses.[3] A GP holds at most one contest in any given school year.

1.2.3. Program costs

We estimate that the average program cost is USD 7.4 per student per year across the two program versions. The variant of the program without GP contests costs USD 6.8 per student; the variant with GP contests costs USD 8 per student. The program's cost thus falls in the middle of the costs of interventions that have improved student learning in India's government schools.[4]

2. Study Design

2.1. 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.[5] Our study includes only "Higher Primary Schools," which end in grades 7 or 8. Before 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

3. 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.

4. One example is a study on a remedial education program that increased math and language learning 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.

5. Appendix Figure A1 shows the study's two districts. In Section 3., 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.

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 before 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 who were enrolled in these sampled schools at baseline.

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 student 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.[6] 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 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).

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

2.2. *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."[7] 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. Appendix Figure A2 depicts the study schools by treatment status. In Section 2.6. below, we check whether the randomization led to comparable groups.

2.3. *Data*

2.3.1. Student achievement

We administered three rounds of standardized math tests to the students in all sampled schools to obtain baseline, midline, and endline assessments. These paper-based tests were administered to students in groups.[8] Assessments had 30-35 multiple-choice items, which are mapped to 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). Test items are also 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. From these previous administrations, we used item response theory (IRT)-based item characteristics to maximize the assessments' test information. Students

7. There was one left-over GP in each district (as 49 is not divisible by four). We paired these two GPs and randomly assigned one to the intervention group and the other to the control group.

8. 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.

had a one-hour time limit; they typically took about 45 minutes to complete each test.

Due to its salience in India, we also administered the "ASER" test of basic arithmetic skill (cf. ASER 2018) to the full sample of students.[9] These tablet-based tests were administered by trained enumerators. The tests are adaptive: They begin with two subtraction problems and, based on a student's performance on these questions, either continue with more difficult (i.e., division) or easier (i.e., number recognition) questions. One-on-one test administration took, at most, ten minutes per student. We followed the ASER's standard grading procedures, classifying 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. We used anchor items across test rounds (baseline, midline, endline) to link all rounds onto a common, continuous ability scale (Kolen and Brennan 2004).[10] 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.

2.3.2. Intermediate outcomes

We collected data 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).

9. 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.

10. 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.

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.[11] 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.[12] In addition, we complement *Teach* with two ancillary data sources for teachers' instructional behaviors: Teacher surveys 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 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.

2.3.3. Implementation fidelity

To capture implementation fidelity in treatment schools, we use two sets of primary and secondary data: 1. Data on teacher training and additional teaching inputs; and 2. Data on 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 surveying teachers on whether they were trained on how to use the teaching and learning materials, the

11. For definitions of each domain and sub-domain, see Molina et al. (2020).
12. Appendix Figure A3 provides mean *Teach* scores, for the control group.

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 community contests data, our 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.

2.3.4. Additional background information

In addition to measuring students' skills, we collected their demographic information to use as additional covariates and to track students 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.

2.4. Timeline

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.[13] 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.

13. 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).

2.5. *Empirical strategy*

We use the following specification to estimate the effects of the program's promotion of activity-based instruction, disentangling it from and comparing it to the intervention variant that also includes community contests.

$$Y_{isgr}^{t} = \alpha_r + \beta_1^t T_{gr} + \beta_2^t D_{gr} + \gamma^t Y_{isgr}^{t=0} + \boldsymbol{\delta'} \boldsymbol{X}_{isgr}^{t=0} + \epsilon_{isgr}^{t} \quad (1)$$

Here, 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 for the program variant without community contests, D_{gr} is a dummy indicating the treatment GPs randomly assigned to contests, 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 β_1^t and β_2^t parameters capture the intent-to-treat (ITT) effect for each program variant, for follow-up round t. We also test whether β_2^t differs from β_1^t.

We also use specifications that allow for heterogeneous treatment effects by interacting potential moderators with the treatment indicator (e.g., student gender).

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. 2023). 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. We describe the statistical methods in more detail in Appendix D.

2.6. *Experiment validity*

2.6.1. Baseline 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 compar-

isons across the three experimental groups in these two tables, 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 (see Ghanem, Hirshleifer, and Ortiz-Becerra 2020). 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.[14]

2.6.2. Implementation fidelity and program take-up

We observed virtually full compliance of GPs' and schools' random assignment to treatment arms, the only exception being 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.[15] 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.

We consider three dimensions of implementation fidelity and program take-up: (i) training, and teacher perception of the program; (ii) teaching inputs and take-up of materials; and (iii)

14. 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.

15. From the random assignment of treatment GPs to community contests (in July 2019), the exposure until endline was approximately 50 percent shorter than the exposure to the activity-based instruction (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 Duflo, Hanna, and Ryan (2012) and Banerjee et al. (2007).

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. Here, we summarize these indicators for the treatment group without contests (see Figure I); implementation fidelity and take-up are similar in the group with contests (see Appendix Figure A5).

For the training and perception dimension, we use both a headmaster and a teacher survey. The headmaster survey shows a high take-up rate: 92 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 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 81 percent of fourth-grade math teachers, with no GKA trainings administered to control-group teachers. Similarly, 87 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 96 percent of the treatment schools at least once, and 75 percent of the treatment schools at least twice, over the study period (they did not visit control schools). Overall, 84 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 (92 percent) treatment-group teachers reported having received the GKA kit. More teachers in treated schools (37 percent) conduct group activities on a daily basis than teachers in control schools (15 percent). Most (59 percent) treatment-group teachers reported using the GKA kit for math classes in every class. Classroom observations using the *Teach* instrument reveal that 40 percent of teachers in treated schools conduct group activities during class. This is a 29 percentage-point difference compared 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 (71 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 (68 percent overall, or 96 percent of the treatment-group teachers who used TLMs).

Finally, we investigate whether these events were implemented as intended. The GP contests took place between August 2019 and January 2020, with 24 days on which contests were held. In Appendix Figure A5, 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 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,[16] although 73 percent of students participated in the contests.

3. Results

Here, we report whether the standalone intervention (without contests) had an impact on intermediate outcomes (especially the pre-registered dimension of instructional quality), whether it improved student learning, and how these effects compare to those in prior studies. We document muted impacts and explore explanations for this finding in Section 4. We investigate whether adding community contests improved the standalone intervention in Section 5.

3.1. Effects on classroom instruction and other intermediate outcomes

3.1.1. Effects on classroom instruction

In Figure II, we present the intervention's effects on teaching quality.[17] We document a positive effect of 0.11 SDs on the overall index of teaching quality, but we cannot conclude with confidence that the coefficient is in fact different from zero ($p = 0.12$). This finding for the overall index masks a positive impact of 0.16 SDs on the pre-specified dimension of teaching behaviors related to activity-based instruction (e.g., whether the teacher promotes student autonomy) and a positive impact of 0.17 SDs on the dimension of teaching that is expected to promote students' socioemotional skills (e.g., whether the teacher promotes collaborative

16. 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 is our research team's observations during the contests.

17. 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.

skills). We do not find statistically significant effects on the dimension of teaching related to classroom culture (e.g., whether the teacher creates a supportive learning environment) and instructional quality (e.g., whether the teacher provides high-quality feedback). In summary, we find positive effects on the index of the three dimensions of teaching related to activity-based instruction and no notable impacts on the remaining sub-domains.

3.1.2. Effects on student attitudes and parental involvement

In Figure II, we document small 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 estimate shows a 0.08 SD improvement, but the coefficient is not significantly different from zero at conventional levels. The standalone program (without community contests) did not target parents and, perhaps unsurprisingly, both parent- and teacher-reported parental involvement are similar to those in control schools.

3.2. Effects on student learning

3.2.1. Average effects

Panel A of Table II summarizes results for the study's main outcome. In the period from baseline to midline, control-school students' math scores improved by 0.13 standard deviations ($p < 0.10$). In the period from baseline to endline, control-school students' math scores improved by 0.40 standard deviations ($p < 0.01$). At midline, 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. At endline, 13 months after the launch of the intervention, we find marginally significant, positive effects (0.12 SDs, $p = 0.11$).

The remaining panels of Table II provide secondary results. First, the program did not lead to notable improvements on the ASER test (see Panel B). Second, the small program impacts at endline are driven by positive effects on lower-order thinking skills (0.14 SDs, $p < 0.05$; see Panel C) and on geometry questions (5 percentage points, $p < 0.01$; see Panel D).

3.2.2. Heterogeneous effects

We further investigate whether the effects on students' 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 III provides the intention-to-treat effects on the study's main outcome measure.

Table III shows that positive program effects are entirely driven by improvements among girls. For girls, we find significant improvements of 0.18 SDs ($p < 0.05$). In contrast, for boys, relative to girls, these effects are 0.15 SDs lower ($p < 0.10$). That is, for boys, coefficients are very close to zero, and they are statistically insignificant. We do not observe clear patterns of heterogeneous effects for the remaining two subgroups of students.

3.2.3. Connecting the results to prior evidence

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 conducted in developed countries 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 III, we compare the findings from the 28 studies identified by this systematic review to our findings. 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 at endline). This occurs even though, at endline, this intervention had provided the longest student exposure to the program. The comparison also shows that previous studies predomi-

nantly 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.

4. Exploring Reasons for Muted Impacts

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 the pre-specified dimensions 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

$$Y_{isg} = \alpha_{sg} + \boldsymbol{\delta}' \boldsymbol{X}_{isg} + \epsilon_{isg} \tag{2}$$

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

$$\lambda_{sg} = \alpha_{sg} - \frac{\sum_{s=1}^{S_g} N_{sg}\alpha_{sg}}{\sum_{s=1}^{S_g} N_{sg}} \tag{3}$$

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

$$V\left(\lambda_{sg}\right) = V\left(\hat{\lambda}_{sg}\right) - E\left\{\frac{\left(\sum_{d=1}^{S_g} N_{dg}\right) - N_{sg}}{N_{sg}\left(\sum_{d=1}^{S_g} N_{dg}\right)}\sigma^2\right\} \tag{4}$$

where σ^2 is the within-school variance of the student-level residual ϵ_{isg}.[18]

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 smaller; we find that a one-standard-deviation increase in school quality (defined with respect to student attitudes towards mathematics) is associated with a 0.19 standard-deviation increase in student attitudes towards mathematics.

Next, we correlate λ_{sg} with teachers and teaching practices (focusing on the index of the three pre-specified subdimension of practices associated with the GKA program). Table IV 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.[19] Column (1) shows that 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 teach-

18. 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)).

19. 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.

ing practices is *negatively* related to student learning (and none of the other subdimensions is distinguishable from zero, at conventional significance levels).[20] 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 ($p = 0.56$).[21]

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, which shows that several dimensions of teacher quality are positively correlated with students' baseline test scores, suggests that this phenomenon exists. Yet, note that better baseline scores correlate with better instruction and greater adoption of activity-based instruction. We expect these schools to be *more* productive than others. 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.

5. Do Community Contests Add Value?

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 (see 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

20. Appendix Table A2 shows the point estimates are similar (albeit slightly more imprecisely estimated) if the value-added estimation focuses on girls' test scores only. Thus, we are unable to link the gender heterogeneity in treatment effects to a differential relationship between activity-based instruction and test score gains for girls (relative to boys).

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

materialize after Round 2. As shown in Table V, 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 A5 does the same for each of the nine underlying *Teach* indicators.

At the same time, the contests failed to increase parental engagement. Perhaps unsurprisingly, as hardly any parents attended these events, Appendix Figure A7 shows that parental engagement remained unaffected. In addition, the effect on student attitudes of the program with contests is very similar to (and statistically indistinguishable from) the effect of the program variant without contests.

While the adverse effects on instructional quality did not translate into reduced learning outcomes, we can rule out that adding the contests led to improvements in student learning. Appendix Table A6 shows that the ITT effects of being assigned to the augmented program variant that includes the contests are close to zero. Coefficients for a comparison with the standalone version of the program are negative (-0.10), and under a directed hypothesis, the 95-confidence interval suggests that we can reject positive effects of 0.06 standard deviations or higher.

6. Robustness Checks

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

Columns (1) to (3) investigate robustness to attrition. We present inverse probability-weighted estimates and Lee (2009) bounds. Columns (4) to (7) investigate robustness to alternative sample definitions. We present results for the study sample of students conditional on participating in any baseline test, the study sample of students conditional on participating in any baseline test and at least a written endline test (but not necessarily an oral endline test), a sample where

we remove a randomization stratum with contamination (one treatment school in the group without community contests received a contest), and a sample of schools that drop the two 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 standalone program 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. Also, 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.

7. Conclusion

Many developing countries have substantially increased their spending on education, which has led to large increases in student enrollment. Nevertheless, 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 perfor-

mance 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 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 large-scale effort to improve teaching and student learning by promoting activity-based instruction. We estimated the causal effects of an innovative program in the state of Karnataka, India, that promotes activity-based instruction 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 teaching 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 isolate the effect of activity-based instruction from a program component that aims to increase community engagement, we randomly assigned the treatment group into two sets of Gram Panchayats, one of which received a version of the program that excluded community contests, while the other received a version that included these contests.

Our analysis shows adherence to this study design: the program was implemented as intended, and that led to the expected changes in pedagogy. More specifically, almost all of the grade-four teachers in program schools received the GKA training, all program schools received the additional teaching inputs (GKA kits), and there were substantial differences in the teaching methods used by the program school teachers (relative to the control group teachers). There-

fore, we consider our study to be a successful test of a "best practice" widely recommended for adoption by education practitioners, outside of its original high-income setting. As in Muralidharan and Singh (2020), we document high implementation fidelity and program adoption; however, we also show impacts on classroom practices. Combined with favorable generalizability conditions—the baseline levels of school productivity and activity-based instruction are low in the study context—our findings thus shed light on whether the potential of activity-based instruction generalizes to developing country settings.

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 promotion of activity-based instruction had small (0.12 standard deviations) impacts on students' mathematics learning. Analysis by gender finds a significant impact of 0.18 standard deviations ($p < 0.05$) 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. Our findings are robust to adding the community-engagement component to the intervention. The estimate for the program variant with the added community-engagement component is close to zero, and we can rule out with precision that the addition of contests led to additional learning gains (added effects of 0.06 SDs or more are ruled out at 95 percent confidence). If anything, we find that the addition of community contests led to a less hospitable classroom environment (-0.40 to -0.49 SDs) in the remaining study period after the contests were conducted).

Our analysis of value-added models provides one explanation for the muted program impacts. We find those schools with greater adherence to activity-based teaching practices exhibited *lower* productivity in raising student test scores. This (non-experimental) finding may explain why the program's impact on instructional practices did not coincide with substantive improvements in test scores. Hence, as governments try to promote activity-based instruction 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. Finally, 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.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* 138, no. 1 (February): 1–35.

Alan, Sule, and Ipek Mumcu. 2022. *Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention.* Working Paper 17601. London, UK: C.E.P.R. Discussion Papers, October.

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.

Angrist, Noam, and Rachael Meager. 2022. *The role of implementation in generalisability: A synthesis of evidence on targeted educational instruction and a new randomised trial.* Working Paper 4. Oxford: Centre for Excellence, Development Impact, and Learning (CEDIL), September.

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.

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

Ashraf, Nava, Abhijit Banerjee, and Vesall Nourani. 2021. "Learning to teach by learning to learn." Cambridge, MA.

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.

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.

Bates, Mary Ann, and Rachel Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review* 15.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

Ministry of Home Affairs. 2012. "15th Census of India."

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

Muralidharan, Karthik, and Abhijeet Singh. 2020. *Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India.* Working Paper 28129. Cambridge, MA: National Bureau of Economic Research, November.

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.

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.

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

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

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.

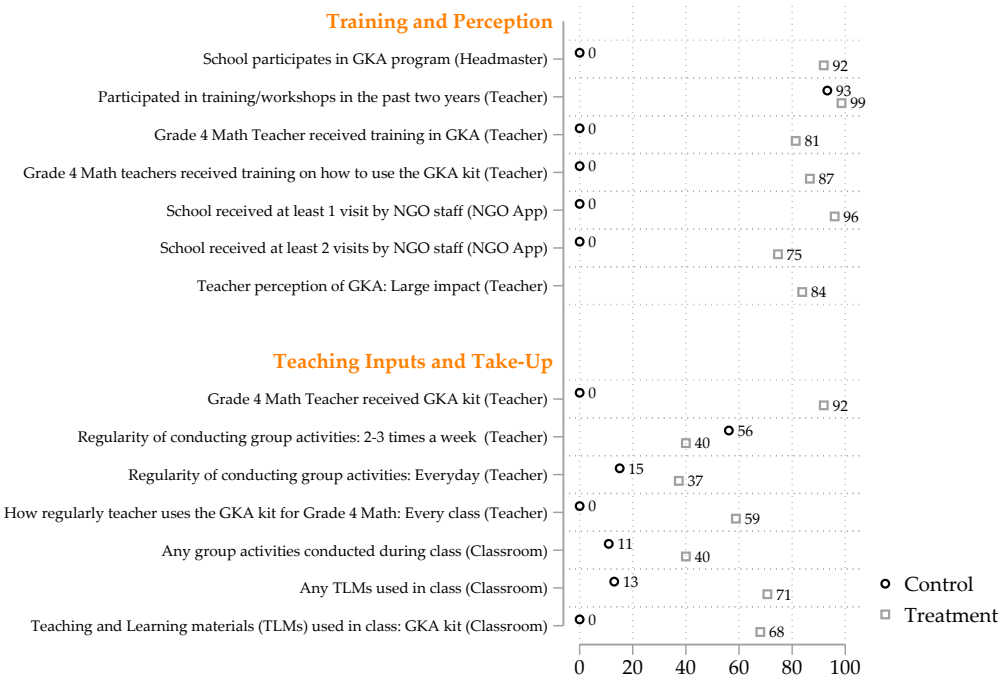
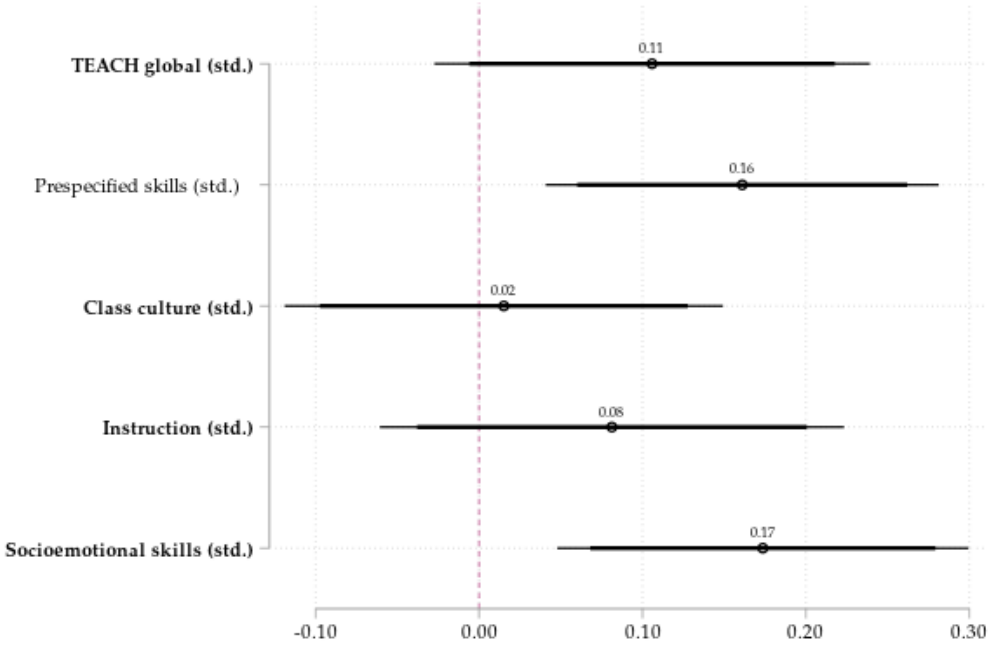
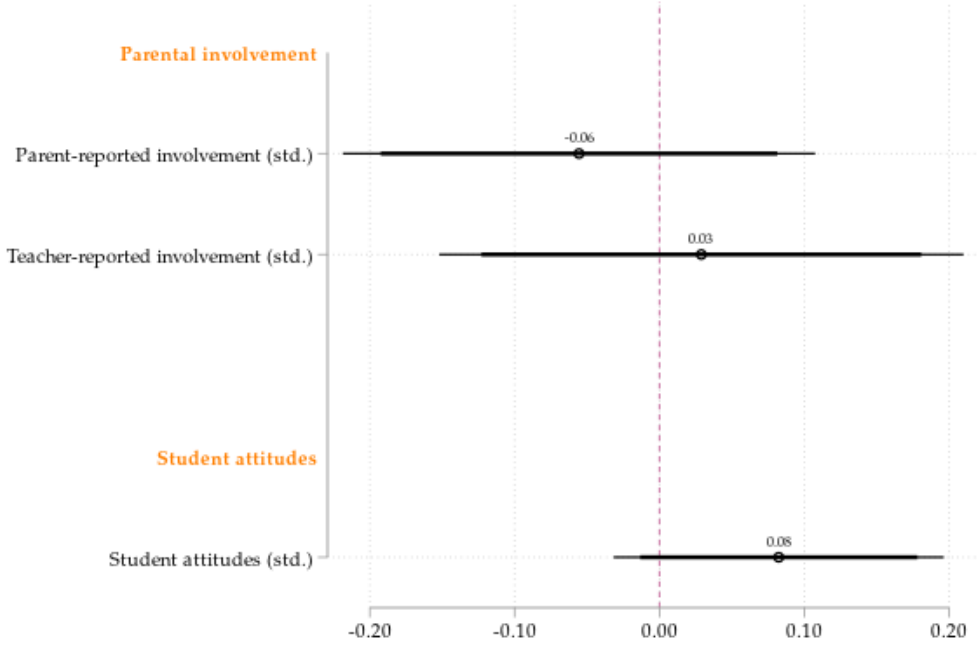


FIGURE I
Implementation fidelity and program take-up

This figure depicts the percentage of teachers or schools that satisfy indicators of implementation fidelity and program take-up, by experimental group. "Treatment" refers to schools in the treatment group without contests.



(a) Effects on classroom instruction



(b) Effects on parental involvement, student attitudes

FIGURE II
ITT effects on instruction, other intermediate outcomes

This figure depicts the program's ITT effects on classroom instruction and other intermediate outcomes. Subfigure (a) depicts the ITT effects on instructional quality. Subfigure (b) depicts the ITT effects on parent self-reported involvement and teacher-reported parental involvement, as well as student attitudes towards math. Thick/thin horizontal bars show 90-/95-percent confidence intervals.

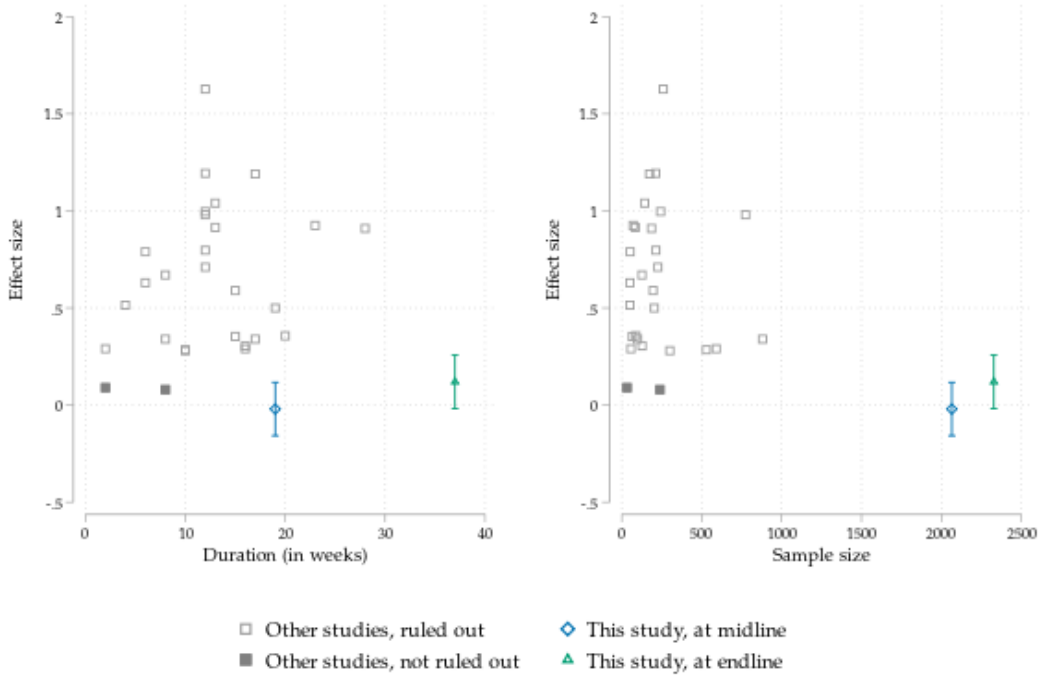


FIGURE III
Comparison with other studies

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 (endline) upper bound. The panel to the left shows the exposure time on the x-axis (in weeks); the panel to the right shows 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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

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***	-0.02	0.12
	[0.97]	(0.07)	(0.08)	(0.07)	(0.07)
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***	-0.02*	-0.02**
	[0.30]	(0.01)	(0.01)	(0.01)	(0.01)
ASER (Baseline)>=Subtraction	0.34	0.09***	0.24***	-0.00	0.02
	[0.47]	(0.02)	(0.01)	(0.03)	(0.03)
ASER (Baseline)=Division	0.09	0.01	0.11***	-0.02	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***	-0.01	0.08
	[0.98]	(0.08)	(0.08)	(0.07)	(0.07)
Lower-order	0.04	0.10	0.31***	-0.01	0.14**
	[0.98]	(0.07)	(0.07)	(0.07)	(0.07)
Panel D: Effects by content domain					
Data	0.38			-0.01	0.02
	[0.21]			(0.02)	(0.02)
Geometry	0.49			0.01	0.05***
	[0.29]			(0.02)	(0.01)
Number sense	0.61			0.00	0.03*
	[0.27]			(0.02)	(0.02)
Whole number operations	0.52			-0.00	0.03
	[0.29]			(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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE III

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***	0.01	0.18**
	[0.98]	(0.07)	(0.08)	(0.07)	(0.09)
Male	-0.05	0.19**	0.51***	-0.06	0.04
	[0.96]	(0.09)	(0.08)	(0.08)	(0.08)
Male vs Female	-0.13*	0.10*	0.20***	-0.07	-0.15*
	(0.07)	(0.05)	(0.06)	(0.06)	(0.08)
Panel B: By baseline learning level					
Bottom tercile	-1.06	0.61***	0.96***	-0.10	0.12
	[0.55]	(0.05)	(0.04)	(0.10)	(0.10)
Middle tercile	0.01	0.20***	0.41***	-0.03	0.06
	[0.25]	(0.04)	(0.04)	(0.09)	(0.08)
Top tercile	1.05	-0.32***	-0.10**	0.07	0.18*
	[0.49]	(0.04)	(0.04)	(0.08)	(0.10)
Top vs bottom tercile	2.07	-0.94***	-1.06***	0.17	0.06
	(1.48)	(0.09)	(0.11)	(0.12)	(0.12)
Panel C: By district					
Bijapur	-0.04	0.10**	0.30***	-0.11	0.16
	[0.76]	(0.05)	(0.04)	(0.09)	(0.11)
Tumkur	0.19	0.18***	0.58***	0.10	0.07
	[0.04]	(0.06)	(0.05)	(0.12)	(0.12)
Tumkur vs Bijapur	0.22*	0.08	0.27**	0.21	-0.09
	(0.13)	(0.14)	(0.14)	(0.15)	(0.17)

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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE IV

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)			
Classroom 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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE V

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)
Classroom 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE VI

ROBUSTNESS OF RESULTS AT ENDLINE

	Attrition			Sample definition				Outcome	Re-randomization
	IPW	Lee (lower)	Lee (upper)	Any BL	Any BL, written EL	No contamination	Complete strata	Written only	Rand. inference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Average effect	0.12*	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 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 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]
Average 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]

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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

Appendix

A Additional Figures and Tables

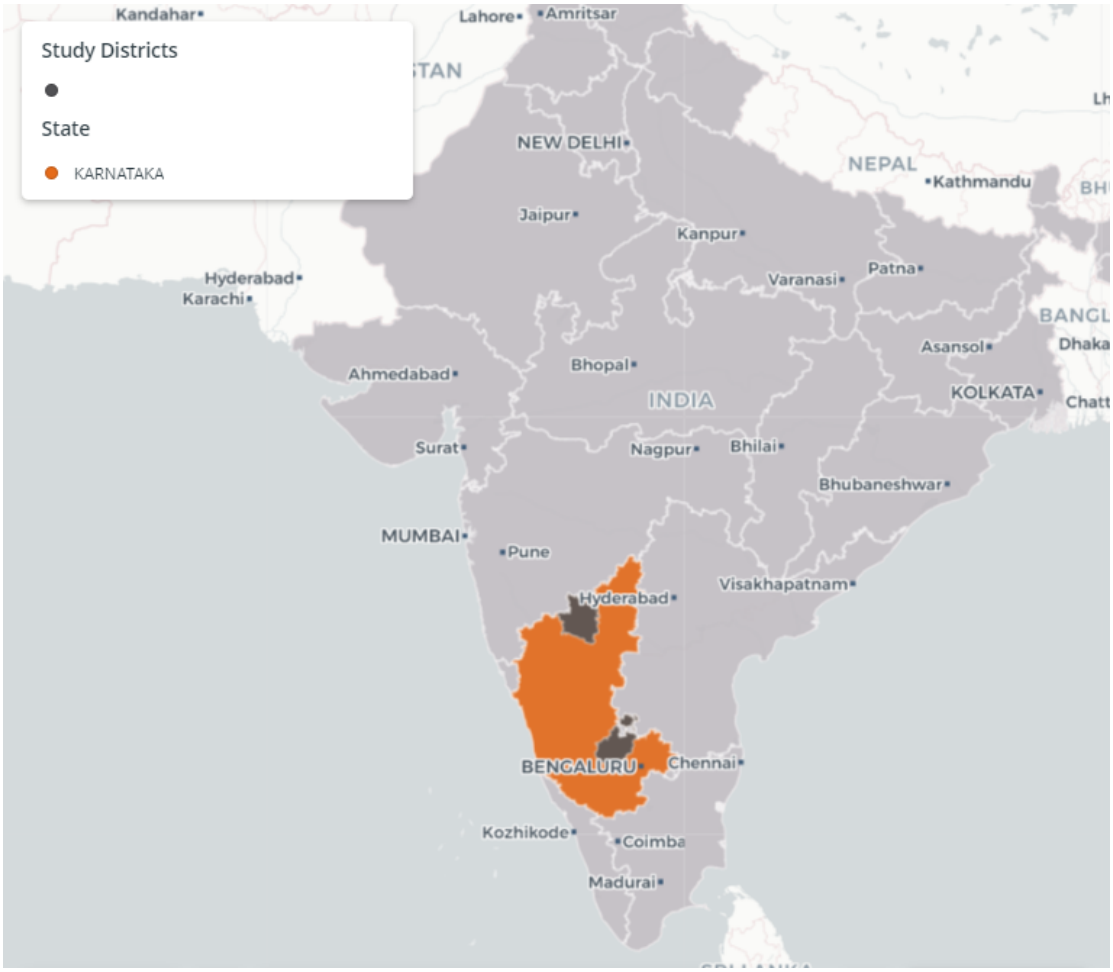
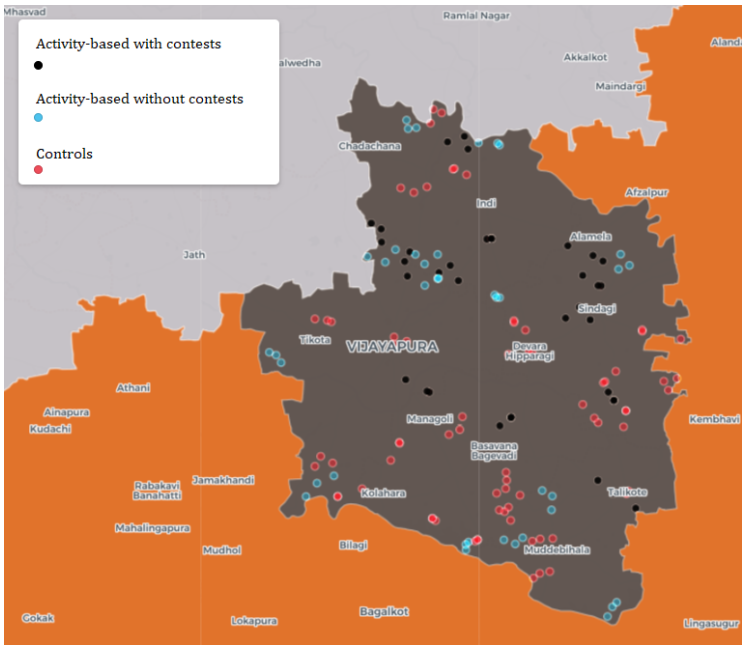
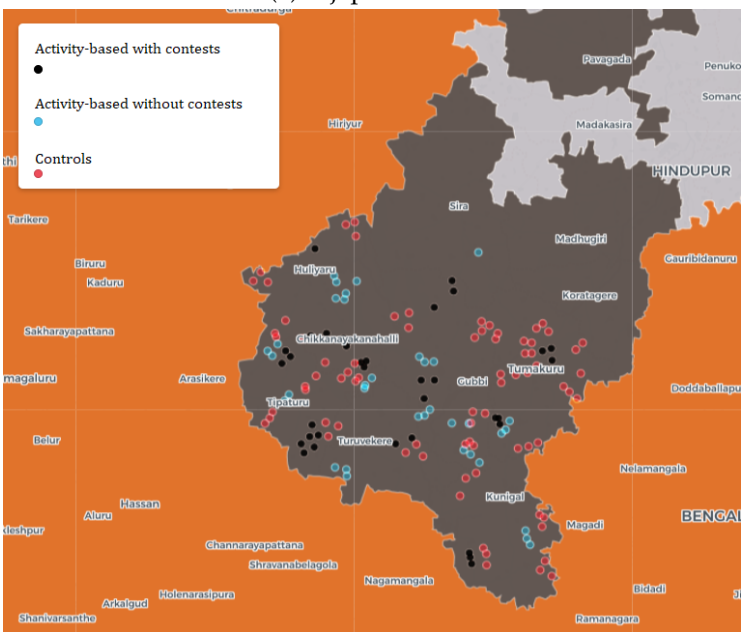


Figure A1
Location of the study

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



(a) Bijapur district



(b) Tumkur district

FIGURE A2
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 the two treatment conditions (with or without contests). Ten re-randomizations to increase balance across T and C, following a "min-max" strategy (cf. Banerjee et al. 2020; Bruhn and McKenzie 2009).

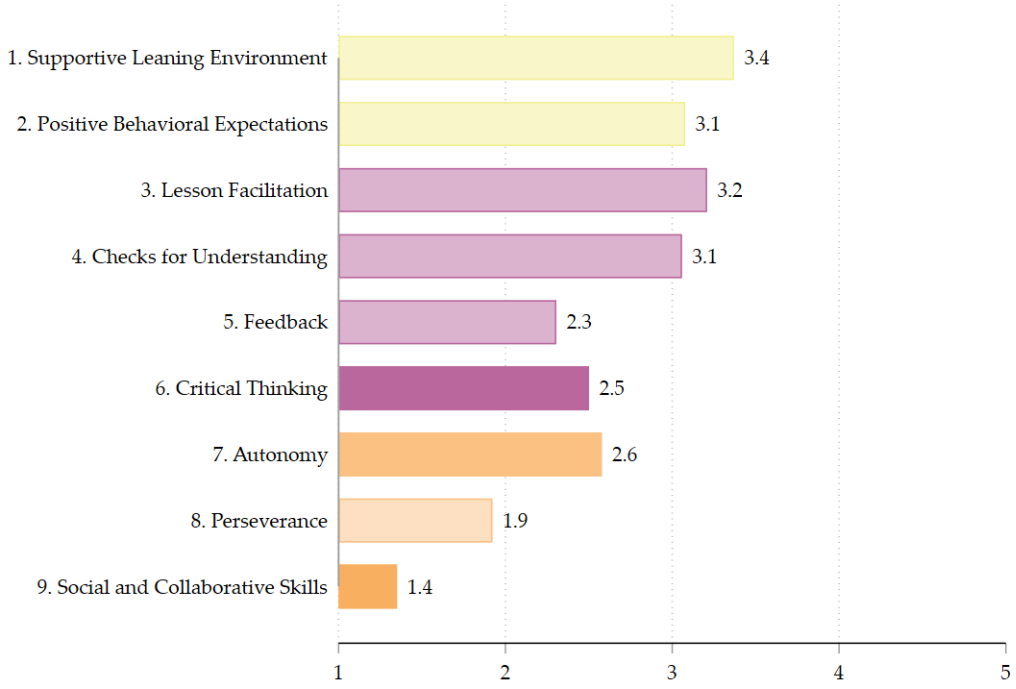


FIGURE A3
Description of teaching quality in the control group

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").

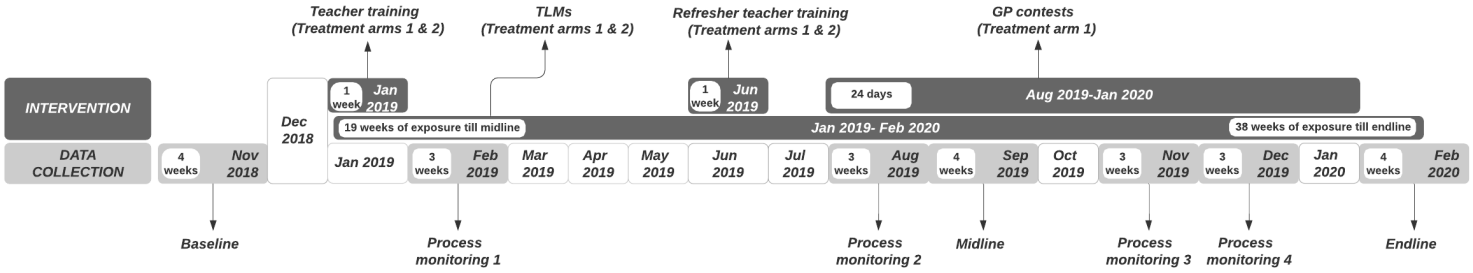


FIGURE A4
Study timeline

This figure depicts the study's timeline. Program implementation activities are shown at the top, in dark gray. Data collection activities are 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. In rounds 1, 2, and 4, "process monitoring" includes student interviews, teacher interviews, and classroom observations. In round 3, "process monitoring," includes parent interviews, teacher interviews, and classroom observations. Not shown: "Data collection" moreover includes (1) observations during all GP contests, (2) administrative records on teacher training events and school visits, and (3) data on program costs.

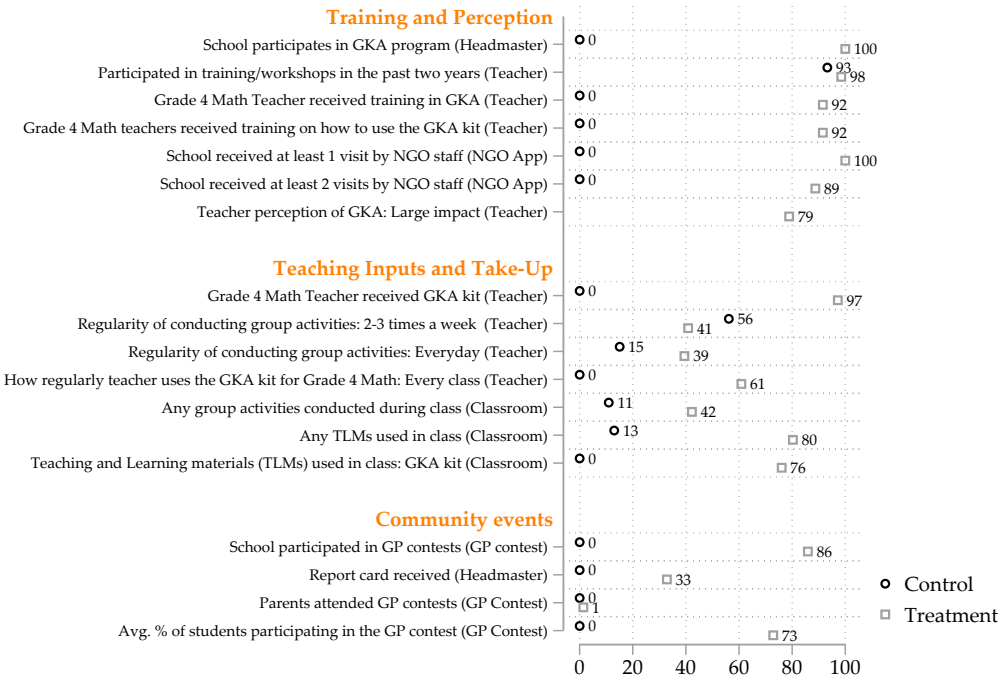


FIGURE A5
Implementation fidelity and program take-up in the group with contests

This figure depicts the percentage of teachers or schools that satisfy indicators of implementation fidelity and program take-up, by experimental group. "Treatment" refers to schools in the treatment group with contests.

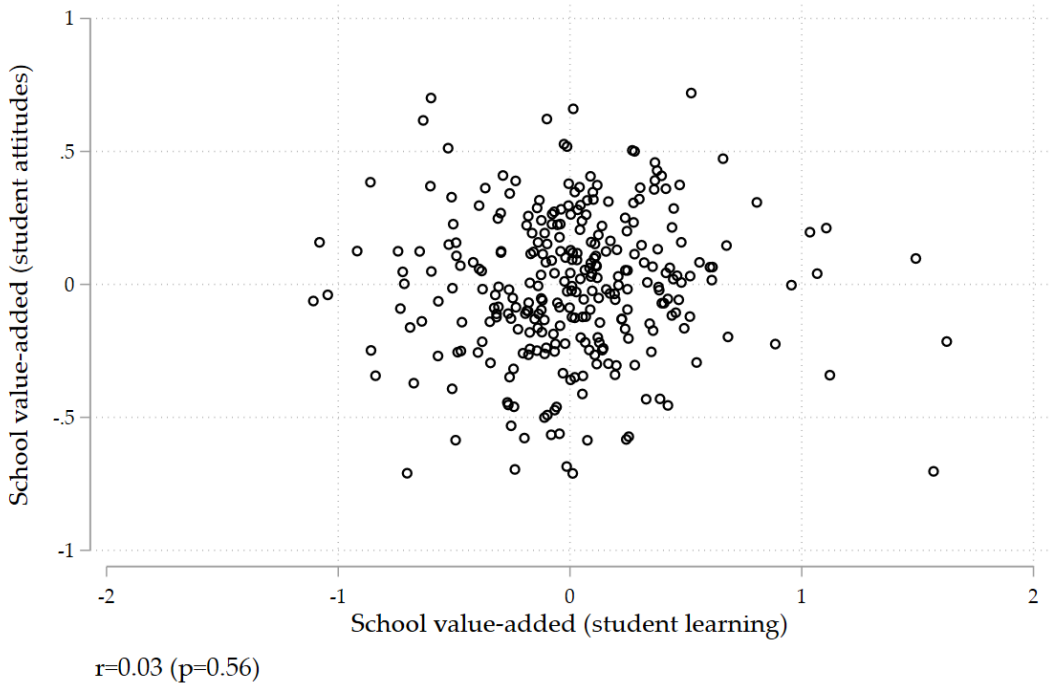


FIGURE A6
Correlation between school value-added on learning vs on student attitudes

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 ($n = 292$).

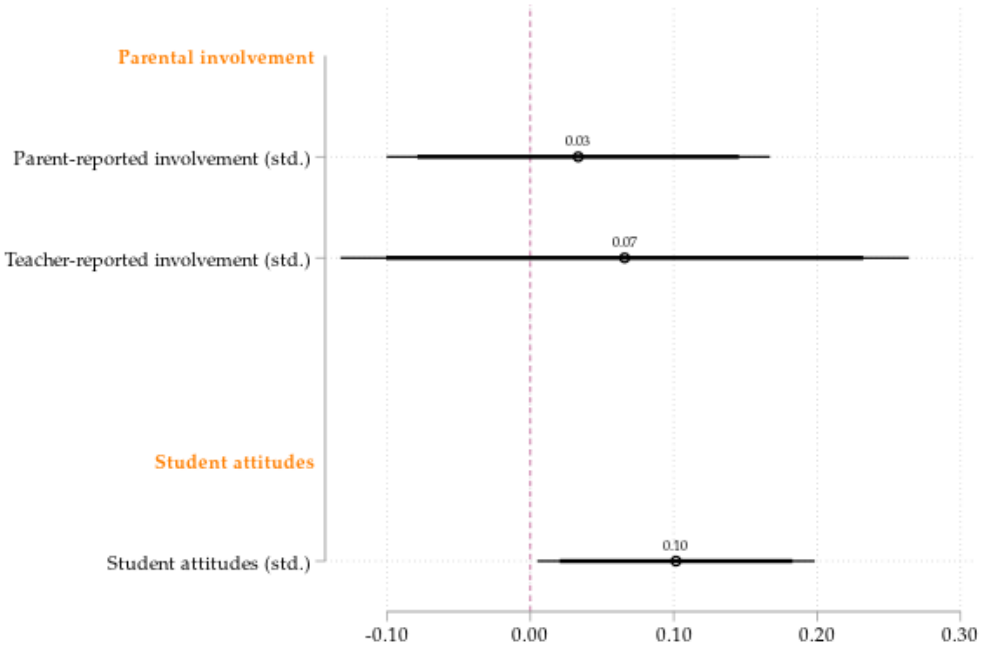


FIGURE A7
ITT effects of program with contests on parental involvement, student attitudes

For the program variant with contests, this figure depicts the ITT effects on parent- and teacher-reported parental involvement, as well as student attitudes towards math. Thick/thin horizontal bars show 90-/95-percent confidence intervals.

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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE A2

School value-added on girls' learning and instructional quality

	(1)	(2)	(3)	(4)	(5)	(6)
Teach global		-0.03				
		(0.03)				
Prespecified skills			-0.05			
			(0.03)			
Classroom culture				-0.02		
				(0.03)		
Instruction					-0.01	
					(0.03)	
Socioemotional skills						-0.04
						(0.03)
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	0.00	0.01	0.01	0.01
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
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	0.08	0.07	0.08	0.08	0.08	0.08
	(0.11)	(0.10)	(0.11)	(0.10)	(0.11)	(0.11)
R^2	0.01	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 female students' 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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

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)			
Classroom 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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

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)			
Classroom 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). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE A5

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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

TABLE A6

ITT effects of adding community contests on student learning

	Control group			ITT effects at midline		ITT effects at endline	
	Baseline mean	Gain to midline	Gain to endline	Contests vs Control	Contests vs Materials	Contests vs Control	Contests vs Materials
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Effects on main outcome							
Main outcome	0.04	0.13*	0.40***	-0.10	-0.09	0.01	-0.10
	[0.97]	(0.07)	(0.08)	(0.07)	(0.08)	(0.09)	(0.10)
Panel B: Effects on ASER test							
ASER>=1-digit	0.99	0.01	0.01***	-0.01**	-0.00	-0.00	-0.00
	[0.09]	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ASER>=2-digit	0.90	0.06***	0.07***	0.00	0.02*	-0.00	0.02*
	[0.30]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ASER>=Subtraction	0.34	0.09***	0.24***	-0.02	-0.02	-0.00	-0.02
	[0.47]	(0.02)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)
ASER=Division	0.09	0.01	0.11***	0.03	0.04**	0.01	-0.01
	[0.29]	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Panel C: Effects by cognitive domain							
Higher-order	0.04	0.08	0.27***	-0.10	-0.09	-0.04	-0.12
	[0.98]	(0.08)	(0.08)	(0.07)	(0.08)	(0.09)	(0.11)
Lower-order	0.04	0.10	0.31***	-0.11	-0.10	0.06	-0.08
	[0.98]	(0.07)	(0.07)	(0.07)	(0.09)	(0.09)	(0.10)
Panel D: Effects by content domain							
Data	0.38			-0.02	-0.01	-0.00	-0.02
	[0.21]			(0.02)	(0.02)	(0.03)	(0.03)
Geometry	0.49			-0.02	-0.04*	0.03*	-0.02
	[0.29]			(0.02)	(0.02)	(0.02)	(0.02)
Number sense	0.61			-0.02	-0.02	-0.00	-0.03
	[0.27]			(0.02)	(0.03)	(0.02)	(0.03)
Whole number operations	0.52			-0.03	-0.03	0.00	-0.03
	[0.29]			(0.02)	(0.03)	(0.03)	(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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.010$.

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.

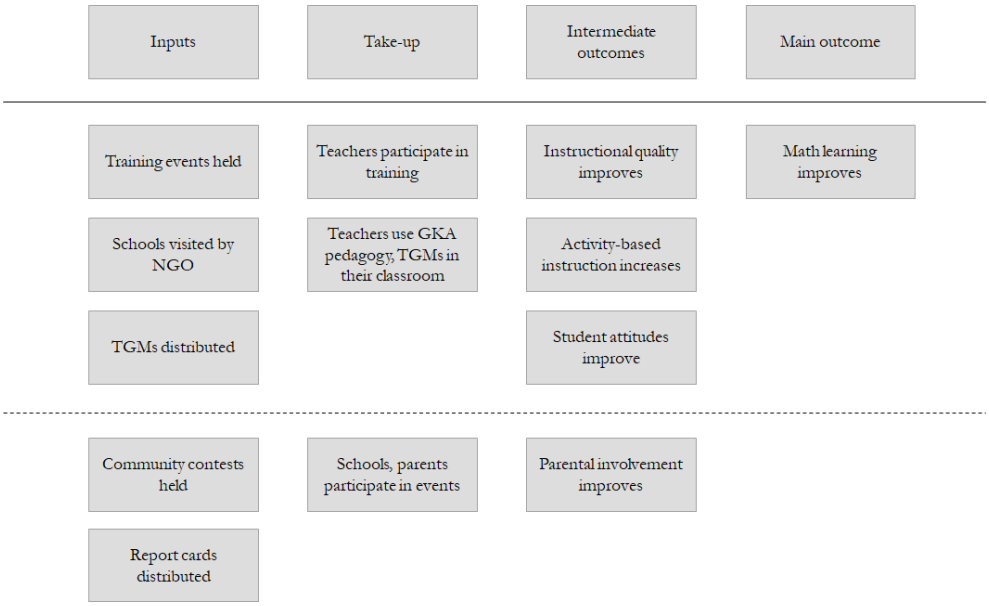


Figure B1
Theory of Change

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.[22] 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).

22. 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.[23] Secondly, for each round of assessments, we reviewed the test questions with an external panel of subject matter experts.[24] 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.

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

24. 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.[25] 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.[26] 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 (Cα) is a widely used measure of reliability in psychometric testing. The Cα is a function of the number of items in a test, the covariance between pairs of items, and the variance of the total score. The theoretical value of Cα varies from 0 to 1, with a rule of thumb of 0.7 or higher suggesting that the test is reliable. In this study, the Cα is 0.91 for the 32 written items. We thus conclude that our instrument is highly reliable overall.

This overall reliability level may nevertheless not translate into high levels of precision for the full range of test takers (as low-ability and high-ability are usually measured with higher levels of noise). Lastly, we therefore consider an additional measure of precision: the test infor-

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

26. 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.

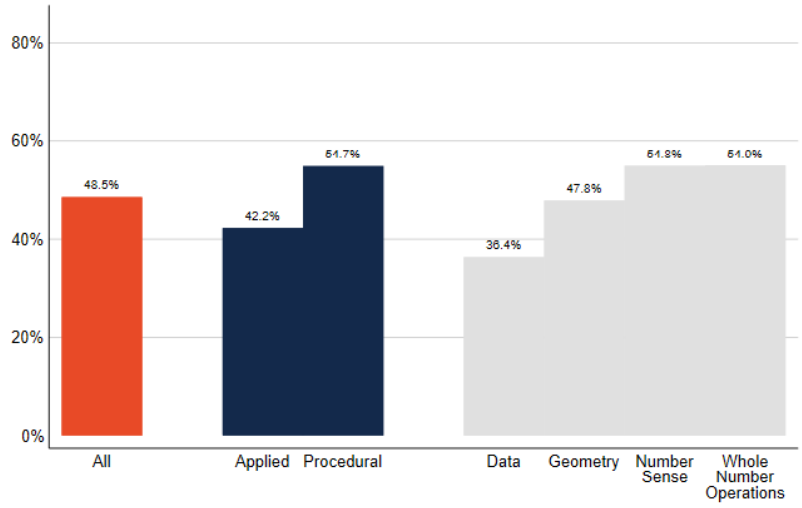


FIGURE C1
Mean percentage of items solved correctly (Baseline)

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-

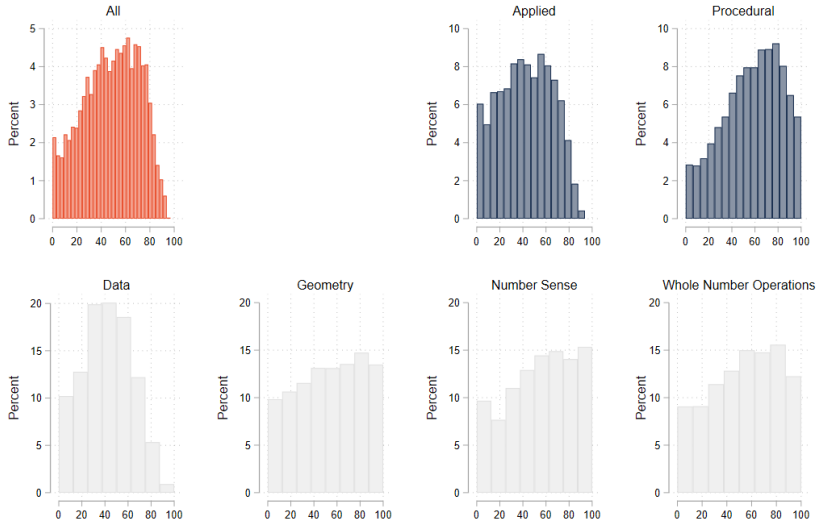


FIGURE C2
Distribution of percentage of items solved correctly (Baseline)

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

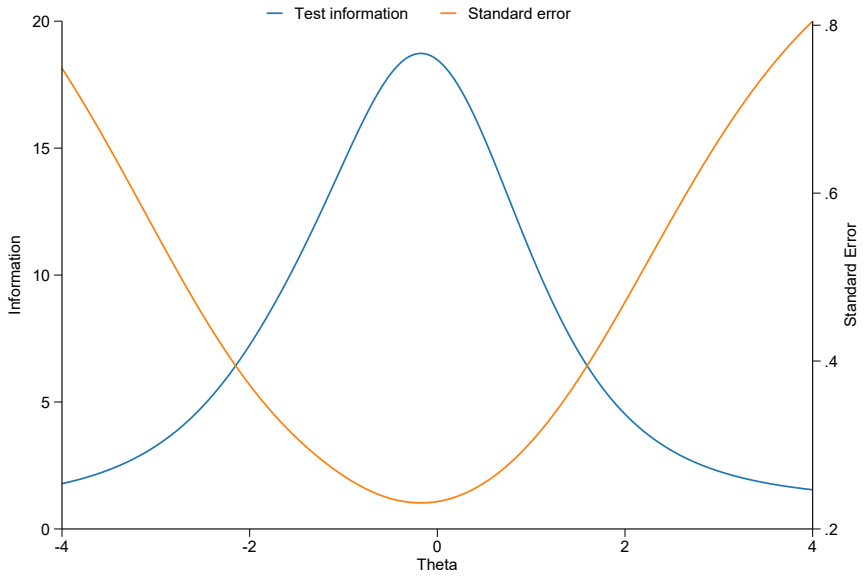


FIGURE C3
Test information function (TIF)

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,[27] and report on program outputs (see Section 2.6.2.).

Spill-overs. We randomized at the GP level; we thus include multiple schools per randomization unit. Therefore, we expect no spillovers 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

27. 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).[28]

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.[29] 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.

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

29. 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).