# Learning by Doing? Experimental Evidence on Activity-based Instruction in India*

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

March 13, 2020

**Abstract**

There are now many rigorous studies on "what works" to increase student learning in developing countries, but there is surprisingly little evidence on how to increase learning by changing instructional practice. We conduct a randomized trial to study the causal effect of an innovative program in Karnataka, India, that promotes activity-based learning through teacher training, additional inputs, and community engagement. For this study, we have randomly assigned 98 administrative units (*Gram Panchayats*) and 292 schools to either receiving the program or a control group. Our primary outcome of interest is child learning, in mathematics, for students in grade four (at baseline). The study's secondary analyses disentangle the effect of individual program components, investigate mediating variables (instructional behaviors, community and parental engagement), and assess the program's implementation fidelity. Sub-group analyses focus on differential effects by students' initial skill level, gender, and geographic location (i.e., district).

**Keywords:** Activity-based instruction; Human capital; India; Mathematics; Primary education.

**JEL codes:** C93; I21; I25.

**Study pre-registration:** Pre-registered at the AEA RCT Trial Registry (AEARCTR-0003494).

---

*This study has been conditionally accepted at the *Journal of Development Economics* (JDE) via pre-results review.

†Ph.D. Candidate, Graduate School of Education and Graduate School of Arts and Sciences, Harvard University. E-mail: adebarros@g.harvard.edu.

‡Research Consultant, Labor Markets Division, Inter-American Development Bank. E-mail: johannaf@iadb.org

§Professor, Department of Applied Economics, University of Minnesota. E-mail: pglewwe@umn.edu.

¶Research Associate, Department of Finance, Carlson School of Management, University of Minnesota. E-mail: sanka010@umn.edu.

# Contents

# 1 Introduction

Many economists agree that higher levels of education increase economic growth (Barro 1991; Hanushek and Woessmann 2008; Krueger and Lindahl 2001; Mankiw et al. 1992; Sala-i Martin et al. 2004). This economic growth will raise incomes and, more generally, increase the quality of life. In the past decade, a large number of countries in the developing world have substantially increased spending on education, which has been followed by increased enrollment in primary education. Government expenditure on education in India has more than doubled between 2006 and 2013 (in constant PPP$; see UNESCO Institute for Statistics 2018). Concomitant with this increased spending, India's primary school enrollment rates have consistently been over 95% for both boys and girls during the past 10 years (ASER 2018). Despite, or perhaps because of, these positive outcomes, policy discussions have recently shifted to the low academic performance of primary school students. For example, only about half of Indian children who are enrolled in *grade five* are able to read a simple paragraph at the *second-grade* level (50.1% of children), or solve a two-digit subtraction problem (52.3% of children) (ASER 2018). These alarming statistics have opened a serious debate on "what works" to improve learning outcomes in this country.

Various studies in the economic literature report on interventions that seem to work to improve learning outcomes in the developing world. In a recent review, Evans and Popova (2016) report that pedagogical interventions and teacher training are among the most effective ways to improve student learning outcomes, as measured by test scores. In particular, large improvements in student learning are observed in pedagogical interventions that match teaching to students' current levels of learning. In another review, Conn (2017) also reports that employing adaptive instruction and teacher coaching techniques are particularly effective for increasing student literacy in Liberia and Kenya. Regarding teacher training on instructional methods, Ganimian and Murnane (2016) find that providing detailed guidance on what, and how, teachers should teach seems effective to increase the skills of low-performing students.

While promising, these conclusions build on a very small evidence base. For instance, de Barros (2018) finds that, out of the 1,754 complete and ongoing trials registered at the AEA registry (501 of which study education), only 16 measure outcomes relating to pedagogy or teaching practices. For language instruction, examples of such studies include Friedman et al. (2010) and Lucas et al. (2014), who emphasize the importance of using new instructional methods and providing accompanying instruction materials to enhance student learning in Sub-Saharan African countries. For mathematics and language, a series of Indian studies point to the promise of allowing teachers to modify their teaching so that it is "at the right level", by re-grouping students and adjusting learning materials for students who lag behind (see Banerjee et al. (2017a), for an overview). Another small body of research points to the potential of training teachers on the job through coaching (Cilliers et al. 2019; Majerowicz and Montero 2018; Popova et al. 2016). Further, a rather

nascent literature seeks to leverage the promise of educational technology to improve instruction (Lai et al. 2015, 2013; Muralidharan et al. 2019).[1] Finally, a recent set of studies seeks to answer the question of why additional teaching inputs often do *not* lead to learning gains, even in otherwise resource-constrained environments. This research asks whether, to be effective, these inputs need to be bundled with complementary interventions (Barrera-Osorio et al. 2018; Mbiti et al. 2019).

Our research adds to this small but important evidence base. In this study, we estimate the causal effects of an innovative, holistic program in Karnataka, India, that promotes activity-based learning of mathematics through additional teaching inputs, related teacher training, and community engagement. This "Ganitha Kalika Andolana" program (GKA) is designed to enable students to learn mathematical concepts and mathematical thinking through games, puzzles and other engaging activities, while allowing them to find creative ways to arrive at a solution—in marked contrast with the conventional chalk-and-talk method commonly used in Indian schools.

Within the economics literature, our paper thus contributes to scholarship on the determinants of learning outcomes, with particular focus on teaching quality and pedagogy, and more specifically on mathematics teaching and pedagogy. Through the use of a novel classroom observation measure, our study recognizes that prior work in this area has largely ignored program effects on teaching effort and instructional quality. We observe the program as implemented as part of a large scale-up, in partnership between the State Education Department and a local NGO ("Akshara Foundation"), in Government schools, with public teachers, during the usual school hours. We thus also aim to add to research on the effectiveness of public programs under government leadership, beyond smaller, tightly controlled pilots (cf. Allcott 2015; Vivalt 2019). Furthermore, our paper provides evidence on the effects of a bundled intervention that seeks "to make inputs work." We therefore also join an important avenue of emerging research that aims to answer why—although quality teaching is considered a key driver for student learning—additional teaching inputs have often failed to produce improvements in cognitive skill.

To estimate the causal effect of this program on student learning in mathematics, we are implementing a Randomized Controlled Trial (RCT). For this, we have assigned 98 administrative units (*Gram Panchayats*[2]) and their 292 schools to either the program or a control group. Our sample of students includes all children enrolled in grade four, at baseline. Our primary outcome of interest is child learning in mathematics, as measured by administering both oral and written mathematics assessments. In secondary analyses we will exploit a separate randomization among treated Gram Panchayats, to disentangle the effect of community events from the remaining intervention. Sub-group analyses will focus on differential effects by students' initial ability level, by gender, and by geographic location (i.e., by district). The study's remaining analyses follow from the program's

[1] A recent review by Escueta et al. (2017) suggests that initially disappointing technology initiatives (such as "One Laptop per Child") focused too heavily on infrastructure, rather than changing pedagogy and instruction.

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

Theory of Change, its program outputs and intermediate outcomes.[3] In terms of *outputs*, we assess whether the program's main components (teacher training, provision of pedagogical materials, community events) are implemented with fidelity. In terms of *intermediate outcomes*, we investigate changes in observed teaching behaviors and instructional quality, by conducting repeat, in-person classroom observations (following a standard rubric). To explore alternative channels, we further interview a random sub-sample of students, parents, and other community stakeholders.

The remainder of this document proceeds as follows. Section 2 describes the proposed experiment, its sampling, and randomization strategy. Section 3 describes the outcomes and hypotheses, and summarizes the measurement methods and data to be used in this evaluation, while Section 4 describes the proposed empirical strategy. Section 5 exploits the study's baseline results to conduct power calculations and balancing tests. Section 6 concludes by providing the timeline of the evaluation.[4]

# 2 Research Design

## 2.1 Intervention

The Ganitha Kalika Andolana (GKA) intervention combines the provision of new instructional materials, related teacher training, and community engagement to improve the mathematics abilities of primary-school students. In this subsection, we briefly describe each of the program's two main components.

The program was initially started with government primary schools in one block of Bangalore Rural District, in 2011.[5] According to its implementer (Akshara Foundation), Karnataka's Government has since committed to scaling the program to all of the state's 44,000 Government primary schools, in a phased manner. Moreover, in 2017 another Indian state, Odisha, has begun to introduce the program, in approximately 4,200 government primary schools.

### 2.1.1 Teaching inputs for activity-based instruction, and related training

The program's first component consists of the provision of additional teaching inputs, and of related teacher training. This component seeks to refocus mathematics instruction on conceptual

[3] As common for Theories of Change, we suggest that, if inputs and activities produce outputs, this should lead to intermediate outcomes, which will ultimately contribute to main outcomes (or goals/impact). For an accessible introduction to Theories of Change, see Gugerty and Karlan (2018, Chapter 3).

[4] Recently, pre-analysis plans and registered reports have led to the introduction of "standard operating procedures", which streamline analytic decisions across experimental research projects (Lin and Green 2016; Lin et al. 2016). Similarly, the presentation of technical decisions in Sections 2, 3, and 4 closely follows that of another pre-analysis plan (de Barros 2018). We thank the Editor in Chief, Andrew Foster, for his guidance with this approach.

[5] The program started with 257 government primary schools, in the Hoskote block.

understanding, rather than learning by rote. Specifically, GKA provides a kit of teaching-learning materials (TLMs) and instructions to teachers, to facilitate activity-based pedagogy. This kit of TLMs includes various items, such as an abacus, a series of shapes, and measuring kits. Each of the materials is mapped to a particular mathematical concept that is required by the state curriculum.

Training is provided to the primary school teachers by a pool of expert teachers that enables them to create activities around the items in the kit of TLMs. In addition to the training provided in the beginning, a field coordinator is appointed at the block level to support the teachers while implementing this new teaching strategy.

#### 2.1.2 Community events

The program's second component consists of community events. These Gram Panchayat Mathematics Contests ("GP contests") convene stakeholders so that they can witness the mathematical performance of school children, during a public assessment. A GP contest starts with a math test for students in the community—they can be enrolled in any school. Following the test, participants discuss the GKA program and additional issues related to education (with an emphasis on children's learning outcomes and the quality of instruction students receive). Thereafter, results of the assessment are announced, the top three students are recognized, and other descriptive performance statistics are presented to community members. Akshara Foundation mobilizes the resources for these contests to be conducted in participating GPs; however, all expenses are paid for by the GP and other local sources. Throughout a given school year, at most one contest is held, per GP.

### 2.2 Sampling and sample

#### 2.2.1 Sampling

Our study is being implemented in two districts in Karnataka: Tumkur and Vijayapura.[6] We purposely selected these two districts to maximize the study's geographic spread and representativeness, within the state. In a first step, we randomly sampled 98 Gram Panchayats (GPs) in these two districts. Within these GPs, we thereafter randomly sampled three schools each, for a total of 294 schools.

Prior to sampling, we declared some schools and GPs ineligible (using administrative data). First, to be able to track students into higher grades, we focus on a particular type of government primary

[6] In India, districts are the largest administrative units within states and territories. There are 30 districts in Karnataka.

schools: "Higher Primary Schools" (HPS).[7] Moreover, the only HPS considered were those with the following three characteristics: (1) the medium of instruction is Kannada (87.5 percent of HPS); (2) the lowest grade is grade four or lower (99.9 percent); and (3) grade four had at least five students in the previous school year (88.9 percent). Finally, for logistical reasons, the sample includes only GPs with at least three eligible schools (84.5 percent of eligible HPS).

Subsequently, the sampling strategy ensured that half of the study's GPs and schools were drawn from each of the two districts. Beginning with a roster of all Gram Panchayats in these districts, our first step was to randomly select 49 Gram Panchayats from each district. This was done using a "probability-proportional-to size" (PPS) technique, where a GP's selection probability reflects its number of eligible schools.[8]

The second step consisted of randomly selecting three schools from each of the 98 GPs. Within each GP, each school was given the same probability of being selected as one of the three schools for the study. Finally, the study includes all fourth-grade students in these sampled schools (as measured at baseline).

### 2.2.2 Variations from the intended sample size

Two schools were removed after baseline, reducing our sample to 292 schools. Baseline data collection revealed no students present in these two schools. Note that these schools were removed *prior* to randomization into treatment and control schools. In Section 4 below, we describe our analytic strategy to address attrition and potential non-compliance (see Section 4.2).

### 2.2.3 Sample and sub-samples

**Sample.** As per the study's baseline data-collection, a total of 5,227 fourth-grade students were formally enrolled in the study's 292 schools. Of those, 4,026 children (77.0 percent) were present during the baseline data collection.[9]

We consider this group of 4,026 students as the study's sample. Our baseline information indicates that, on average, these students were about 9 years and 2 months old. Approximately 53.0 percent of the sample is female.[10]

[7] The majority (70.4 percent) of Higher Primary Schools end with grade seven; about a quarter (25.1 percent) end with grade eight. About half (45.6 percent) of Karnataka's government primary schools are HPS; the remainder are "Lower Primary Schools" (LPS), which serve grades one to five.

[8] Given their large size, three GPs were included as certainty units (two in Tumkur and one in Vijayapura). The remaining GPs were selected using PPS.

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

[10] Here, we report students' average age and gender as approximate numbers. This is since, for a small percentage of students (2.0 percent), this information is missing. in our follow-up rounds of data collection we will attempt to collect these data.

In Section 5 below, we use the study's baseline data for power calculations, in order to determine the study's minimal detectable effect size (MDE) (see Section 5.1).

**Sub-samples.** For our analyses of intermediate outcomes, we conduct interviews with sub-samples of students and parents. For this purpose, we randomly select (up to) eight students per school, using the baseline roster and students' performance on the baseline test.

More specifically, we stratify each school's list of students by (a) gender (female/male) and (b) baseline performance (above/below the school median). We then select two students per stratum, at random.[11] We follow the same procedure to generate a separate sub-sample of parents; however, for logistical and budgetary reasons, we only select four parents per school. We repeat these sub-sampling procedures separately, for each surveying round.

## 2.3 Randomization

### 2.3.1 Randomization of treatment and control units

To increase statistical power and to warrant balance across treatment and control units, we conducted a stratified randomization to assign the 292 schools to be treatment or control schools. After the baseline test, within each district we used baseline test scores to create quadruplets of Gram Panchayats with similar academic performance.[12] Thereafter, for each of these strata, two GPs were randomly selected to participate in the GKA program, while the other two GPs remained as "controls."[13] Thus, 49 GPs and their selected schools were assigned to receive the program; the remaining 49 GPs and their selected schools continue with "business-as-usual."[14]

We repeated the above-mentioned randomization procedure ten times, to select the randomization with greatest balance. To this end, we selected a vector of covariates—from India's District Information System for Education (DISE)—that are predictive of baseline scores. Thereafter, we calculated *t*-statistics for each of the selected variables across the two groups of GPs, as well as the baseline math score. We did so by estimating regressions of each characteristic on the treatment

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

[12] Specifically, we calculated the average performance score among all students in a given GP, from the baseline's one-on-one test (see below). For logistical reasons, we were not able to use the paper-based section of the test to stratify GPs.

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

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

indicator and strata fixed effects. We then stored away the most extreme of these $t$-statistics, and selected the randomization where this value is smallest.[15]

In Section 5 below, we use the study's baseline data to investigate whether this randomization strategy led to comparable groups (see Section 5.2).

#### 2.3.2 Randomization of community events among treatment units

In addition to the randomization strategy described in 2.3.1, after selecting the randomization of schools with the greatest balance, we randomized all of the 49 treatment pairs into two arms: one group of GPs with community events (24), and one group of GPs without the community events (25). Both treatment arms continue to get the kits and related training. All pairs of control GPs remained untouched. This randomization for the GP contests took place in July, 2019.

# 3 Hypotheses, Outcomes, and Data

## 3.1 Primary hypothesis and main outcome of interest

The study's main research hypothesis is that the program improves student learning, in mathematics. We measure this main outcome of interest in two ways: (1) Student math scores on standardized tests, and (2) Student math scores on a one-on-one test that captures basic mathematical skills.

#### 3.1.1 Standardized math tests

Standardized math tests are administered to the students, in all sampled schools, as a baseline, a midline, and an endline assessment. These tests are paper-based and administered to students as a group.[16] Assessments have 30-35 multiple-choice type questions and students receive a one-hour time limit.[17]

[15] See Bruhn and McKenzie (2009), who refer to this approach as "minmax method." We are well aware that high numbers of re-randomization can lead to analytic problems, especially if the re-randomization strategy remains unknown. We follow Banerjee et al. (2017b) by pre-specifing our strategy and choosing a conservative number of re-randomizations (ten re-randomizations).

[16] During baseline, we were concerned that weak students may not be able to answer a paper-based test. Therefore, we administered a subset of seven items both orally (one-on-one), and as written items. We did not find floor effects and found our concerns to be unwarranted. Results are available upon request. In subsequent rounds, our standardized tests will therefore rely on written (group) administration only.

[17] During baseline, students commonly took approximately 45 minutes to complete their test.

Test items are mapped to the official state curriculum, but also include items from up to two years below grade-level. These items have been administered in similar contexts previously, in India, in large-scale assessments. The assessments do not draw questions from Akshara's internal item bank.

From these previous test administrations, we use item response theory (IRT)-based item characteristics to maximize the assessments' test information.[18] Estimates of student ability are calculated using a standard, three-parameter logistic (3PL) IRT model, with a single guessing parameter (Birnbaum 1968; Samejima 1973).[19] In doing so, anchor items across test-occasions (baseline, midline, endline) will allow for the linking of estimates onto one common, continuous ability scale (Kolen and Brennan 2004; Stocking and Lord 1983).

We provide a more detailed description of the baseline test design and related validity evidence in Appendix 8.2. Our analyses presented in Appendix 8.2 confirm that the test did not suffer from floor or ceiling effects. These analyses also suggest that our test questions discriminate well; that is, students with low ability have a much smaller probability of correctly responding a difficult question than students of higher ability. The test moreover exhibits low levels of noise, both in terms of its overall reliability (as per Cronbachs's alpha) and in terms of its precision for a wide range of test takers with differing levels of ability (as per its test information function).

#### 3.1.2 One-on-one tests of basic mathematical skills

Due to its salience among policy makers, we also administer a well-known one-on-one test of basic mathematical skill (at the same time as the written assessments). These "ASER" math tests (cf. ASER 2017)[20] are administered to our full sample of students. Tests are tablet-based and administered by trained enumerators. One-on-one test administration takes a maximum of ten minutes, per student. We follow ASER's standard grading procedures, which group test takers along five progressive (and mutually-exclusive) levels of ability: beginner, recognition of single-digit numbers, recognition of two-digit numbers, two-digit subtraction (with borrowing), and three-digit by one-digit division (cf. ASER 2017).

### 3.2 Secondary hypotheses and related outcomes

We specify three sets of secondary hypotheses and describe their respective outcomes here, along with the program's Theory of Change. First, we investigate the program's *impact* on student learning along more fine-grained sub-competencies of mathematical skill. Second, we assess three areas of *intermediate outcomes*: whether the program improved instructional behaviors, whether the program

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

[19] In case a 3PL model does not converge, a 2PL model will be used instead.

[20] The ASER is a comprehensive household survey of rural India, which measures the enrolment status and tests basic reading and arithmetic abilities through a common set of testing tools, for children between 3-16 years.

changed students' attitudes towards mathematics, and whether the program increased community engagement and parental involvement. Third, we assess the program's implementation fidelity and its immediate *outputs*.

### 3.2.1 Measures of sub-competencies

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

The tests capture students' ability on four content domains:

1. Basic number sense / number concepts;
2. Whole number operations;
3. Shapes and geometry;
4. Data display, measurement, and statistics.

Each of the test questions is mapped to one of these content domains. The tests also capture students' ability on two cognitive domains:

1. Knowing;
2. Reasoning and applying.

Each of the test questions is mapped to one of these two cognitive domains.

To construct summary outcome measures for each of the six domains, we will calculate the percentage of related test questions a student was able to solve correctly.

### 3.2.2 Intermediate outcomes

**Measures of instructional behaviors.** We will measure time-on-task, instructional quality, and instructional behaviors in treatment and control schools, after the implementation of the program. We will collect these data through unannounced classroom observations. In scheduling these classroom visits, we will follow the study's sample of students—not a given mathematics teacher. Thus, we maintain a strict focus on the instruction these students actually receive, regardless of whether their teachers change over time. We have already conducted a first round of classroom observations in the first school year (June 2018 to May 2019), and we will conduct three additional rounds in the second school year (June 2019 to May 2020).

More specifically, we will use a novel, standardized classroom observation instrument, developed by the World Bank, called "Teach". We selected this instrument for its relevance to the program's Theory of Change, for the academic rigor that went into assessing its psychometric properties, and since it was constructed in (and not merely transferred to) developing countries. We adhere to the instrument as closely as possible; however, the tool has been further piloted in and contextualized for government schools in Karnataka. Teach is organized into three broad domains of instructional quality: Classroom Culture, Instruction, and Socio-emotional Skills; each domain is clearly mapped to respective behavioral markers. The instrument covers a total of nine narrower sub-domains, of which we expect to find higher impacts among the following three dimensions: (1) Critical Thinking, (2) Autonomy, and (3) Social and Collaborative Skills. In addition to instructional quality, the observation tool captures time-on-task, through timed snapshots. To construct summary outcome measures (for domains and sub-domains), we follow Teach's standard procedures, as documented by Molina et al. (2018).

We consider Teach our *main* measure of instructional behavior. However, we will complement this information with two additional data sources: Teacher surveys (during school visits) and student surveys (with the sub-sample of interviewed students). In particular, we focus on teachers' self-reported awareness of activity-based teaching methods, and their use of collaborative pedagogy. During student surveys, we also ask three questions that are intended to measure student-reported quality of instruction.[21] We moreover ask students three questions about their level of collaboration with peers.[22] We will generate summary indices from these items, by calculating inverse covariance-matrix-weighted averages (following Anderson 2008).

**Measures of student attitudes towards mathematics.** Using surveys with the sub-sample of interviewed students, we measure children's attitudes towards mathematics learning. To this end, we administer a battery of four questions.[23] We will generate a summary index from these four items, by calculating their inverse covariance-matrix-weighted average (following Anderson 2008).

**Measures of parental involvement and community engagement.** During student interviews, we administer a battery of questions on parental involvement in their math education. During teacher interviews we will elicit teachers' perceptions on parental involvement, including the last time they communicated with a parent. In our interviews with the sub-sample of parents we also seek to measure parents' involvement in their child's math education.

[21] We ask (a) whether students have difficulty understanding explanations, (b) whether the teacher provides interesting tasks during class, and (c) whether the teacher explains concepts again, if there are doubts.

[22] We ask (a) whether students ask classmates for help, (b) the extent of student collaboration during math class, and (c) about their level of collaboration on homework.

[23] We measure (a) whether the student enjoys learning math, (b) whether math makes the student nervous (c) whether the student finds math hard to understand, and (d) whether the student finds math harder than other subjects.

To gather information on community engagement, we further ask each school's headmaster about activities of their school's School Development and Monitoring Committee (SDMC), as well as about meetings between parents and teachers.[24] During our process monitoring rounds of the second school year, we moreover complement these data by interviewing a school's Gram Panchayat leader and block education officer (BEO).[25]

### 3.2.3 Implementation fidelity and program outputs

We will use primary and secondary data to track implementation fidelity in treatment schools.[26] We organize these data sources along with the program's two main components: Teacher training and additional teaching inputs, and community events ("GP contests"). In the following subsections, we describe these measures in greater detail; Table 1 provides an overview.

[Insert Table 1 here.]

**Teaching inputs for activity-based instruction, and related training.** We collect administrative records on teachers' participation in GKA training sessions, from the Akshara Foundation. We complement these data by asking all teachers about their participation in and perception of these trainings.

During school visits, we record teachers' self-reports on the availability of, and their use of, GKA materials. As part of the teacher survey, we collect data on whether the teachers were trained on how to use the teaching and learning materials, availability and usage of the materials, and their perceptions on the program. We further complement this information with classroom observations and school surveys of the availability and use of teaching and learning materials provided by the Akshara Foundation.

Finally, we intend to gather administrative information on Akshara Foundation's monitoring efforts and (on-site) teacher re-trainings. The Akshara Foundation requires its field staff to document any school visits through a mobile app. We will have access to this information and we will thus be able to collate the number of school visits, per school.

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

[25] BEOs oversee the provision of primary and secondary education within a block. BEOs are responsible for a wide range of tasks, including human resource management, school inspections and monitoring, academic support, and community engagement. For additional information on Block Education Officers, including a list of their tasks, see Aiyar and Bhattacharya (2016).

[26] See Sabarwal et al. (2014), on the importance of measuring program take-up thoroughly.

**Community events.** The research team will attend all community events ("GP contests"). During these events, we will generate student attendance rosters, which will be mapped to the study's sample of children (by use of unique student IDs). At each contest, the research team will also count the number of parents who attended.

We will also record measures of the community engagement through the surveys of Block Education Officers (BEO) and students. Specifically, we will collect data on the BEO's knowledge and perception of the GKA program. In addition, we will ask the students if they have participated in the GP contests.

## 3.3 Disentangling the effect of program components

It appears important to disentangle the effect of community events from the effect of the remaining program. For example, a recent learning-by-play intervention in Ghana led to positive results only if parents were *not* involved (Wolf et al. 2019). Our secondary work therefore investigates treatment effects separately, depending on whether the program includes the community event component, or not.

## 3.4 Cost data

We will collect data on program implementation costs (planned and actual) from the Akshara Foundation.

To estimate the cost-effectiveness of the activity-based learning of the GKA mathematics program to increase students' math test scores, we start by collecting information on basic demographics. We use the number of students in each group (treatment and control) so that we can later multiply the costs and impacts of the program. This information will allow us to calculate the baseline number of students who are eligible for the program and the number of students chosen to participate in the program.

Next, because we will include the cost of beneficiary time to participate in the program, we will estimate the time that families spend on different aspects of the program (e.g., time to travel to and attend community meetings). We also calculate the opportunity cost of parental time, as measured by average daily wages in the study area.

In order to convert all cost and impact calculations to their present value in USD for a standard year of analysis, we must use a few outside pieces of information. More specifically, we will need PPP exchange rates and annual GDP deflator inflation rates for the years 2018-2020. For this analysis, we assume a discount rate of 12%, as suggested in Dhaliwal et al. (2013), using the social opportunity cost of capital (SOC) approach. Our base year will be 2018.

## 3.5 Additional covariates

The following demographic information is collected from students, for use as additional covariates and to facilitate the tracking of students over the study's multiple rounds of data collection: (1) Name; (2) Gender; (3) Birth date; (4) State-issued ID number; (5) Father's name; and (6) Mother's name.

Moreover, we collected additional administrative information for each school (at baseline). To this end, we can leverage data from official school report cards as per the District Information System for Education ("DISE"), as well as information on a school's village, from India's 2011 Census.[27]

## 3.6 Data collection and processing

We are adhering to J-PAL South Asia's strict data collection procedures, including double-entry of paper-based tests, high-frequency checks for electronic forms, spot-checks and accompaniments, and weekly monitoring and de-briefs for field staff (see Glennerster 2017; J-PAL 2017). In the Appendix, we use our baseline data collection to demonstrate how these quality control mechanisms are implemented in practice (see Section 8.1).

# 4 Empirical Analysis

## 4.1 Statistical model

### 4.1.1 Average effects

We will estimate the impact of the program on outcomes, using the following specification:

$$Y_{isg}^{t} = \alpha_g + \beta^t T_{isg} + \gamma^t Y_{isg}^{t=0} + \delta' X_{isg}^{t=0} + \epsilon_{isg}^{t} \tag{1}$$

where $Y_i$ is the outcome of interest for student $i$ in school $s$ and GP $g$, at time $t$. In our primary analysis, $Y_i$ refers to test scores. In our secondary analyses, $Y_i$ will consist of: (a) measures of sub-competencies; and (b) mediating variables. The $\alpha_g$ parameters are randomization strata fixed effects, $T_{isg}$ is the treatment dummy and $\epsilon_{isg}^{t}$ is the idiosyncratic error term. To increase precision, all specifications include $Y_{isg}^{t=0}$ and $X_{isg}^{t=0}$ as covariates. Measured at baseline ($t = 0$), $Y_{isg}^{t=0}$ refers to a student's initial outcome of interest; $X_{isg}^{t=0}$ refers to a vector of baseline controls selected through a LASSO procedure, from student age, gender, school-level DISE data, and village-level census data

[27] We will use GIS to match each school's geographic location to its closest village.

(cf. Dhar et al. 2018). The coefficient of interest is $\beta^t$, which captures the program's intent-to-treat (ITT) effect for the intervention, for each follow-up round $t$.

#### 4.1.2 Effects by program component

In secondary analyses we will investigate the additional effect of community events through the following specification:

$$Y_{isg}^t = \alpha_g + \beta_1^t T_{isg} + \beta_2^t D_{isg} + \gamma^t Y_{isg}^{t=0} + \delta' X_{isg}^{t=0} + \epsilon_{isg}^t \quad (2)$$

where $D_{isg}$ is a dummy indicating treatment GPs that have been assigned to community events, and all else is as in Equation 1. $\beta_2^t$ thus indicates whether treatment effects are equal without or with the events. We will test for whether $\beta_1^t$ alone or the sum of the two coefficients $\beta_1^t$ and $\beta_2^t$ is zero (capturing the ITT effect of the program without and with the community events, respectively).

#### 4.1.3 Heterogeneous effects

We will also use a specification that allows for heterogeneous treatment effects, by interacting potential moderators with the treatment indicator. We illustrate the corresponding specification for the sub-group analysis by gender as follows:

$$Y_{isg}^t = \alpha_g + \beta^t T_{isg} + \theta^t T_{isg} * F_{isg}^{t=0} + \zeta^t F_{isg}^{t=0} + \gamma^t Y_{isg}^{t=0} + \delta' X_{isg}^{t=0} + \epsilon_{isg}^t \quad (3)$$

Here, $F_{igs}^{t=0}$ is the moderating variable of interest (in our illustration, an indicator for a student's gender), measured at baseline, and all else is defined as above.

To avoid specification searching, we limit our analyses of heterogeneous effects to the following moderators:

1. Gender (as illustrated above);
2. Initial level of ability;
3. District.

## 4.2 Statistical methods

### 4.2.1 Estimation

Our default is to estimate standard OLS regressions; in the case of ASER data (which provides five, ordered outcome categories), we will use an ordered logit model. We will cluster standard errors at the GP level (cf. Abadie et al. 2017).

In robustness checks, we will use randomization inference to assess whether the re-randomization procedure led to unexpected consequences (Young 2019). More specifically, we will replicate the same re-randomization procedure within each of 1,000 iterations (cf. Heß 2017).

### 4.2.2 Non-compliance

**Lack of take-up.** Schools and teachers may not take up the treatment. We posit that the policy-relevant question is whether the program led to learning gains even in the light of (potentially) diluted treatment exposure. Our study will therefore analyze intent-to-treat (ITT) effects. However, we will also investigate the effectively observed program exposure[28], and report on program outputs (see Section 3.2).

**Spill-overs.** We have randomized at the GP level; we thus include multiple schools per randomization unit. Therefore, we do not anticipate spill-overs from treatment to control schools. Yet, our school visits will track schools' potential exposure to other, similar interventions (in both groups of schools). In particular, a similar program has been implemented in Karnataka, promoting activity-based instruction in the lower grades ("*Nali Kali*"). There is no overlap between this program and the grade-levels investigated in our research; nevertheless, we will track the potential use of "Nali Kali" pedagogy in treatment and control schools, through the study's classroom observations.

### 4.2.3 Missing values and attrition

We foresee two types of missing values, as follows. Observations may either be observed with incomplete data, or observations may not be observed at a follow-up round of data-collection ("attrition").

[28] In the experimental literature, "exposure" and "dosage" are sometimes used interchangeably. Here, we prefer the term "observed exposure" to more clearly distinguish subjects' effectively experienced levels of treatment from their initially intended levels of treatment.

**Missing data for observed observations.** Students may leave individual test questions blank. We will classify unanswered questions as incorrect answers.

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

We will not impute values to covariates for observations with missing values (e.g., through mean imputation, or multiple imputation). Instead, we will investigate the robustness of results under: (a) the inclusion of covariates, dropping those observations where covariates have missing values ("listwise deletion") vs. (b) the exclusion of those covariates that have missing values.

**Attrition.** We plan to investigate attrition in three ways. First, we will investigate whether attrition is systematically related to treatment status, through tests of differential attrition rates and tests of selective attrition.[29] Second, we will employ robustness checks. More specifically, we will consider inverse-probability weighting (IPW) and the use of Lee (2009) bounds. Third, in the unlikely event that entire schools attrit, we will investigate the robustness of results to dropping those schools from the sample that share the same randomization stratum.

#### 4.2.4 Outliers

All of our outcome variables fall within a limited, pre-defined range. For example, a student may solve all (or none) of the test questions correctly. We do not investigate effects on self-reported outcomes that may include outliers (such as income).

In contrast, covariates may include outliers. In this case, we will consider transformations of covariates and the robustness of results to the inclusion/exclusion of covariates. We will not trim observations; we also will not winsorize observations.

#### 4.2.5 Multiple outcome and multiple hypothesis testing

We account for multiple hypothesis testing by clearly pre-specifying two measures of student learning (paper-based tests, ASER tests) as primary outcomes of interest. We interpret these as "family" measures of mathematical ability, as in similar approaches that use summary indices to counter the issue of multiple hypothesis testing (Anderson 2008; Kling et al. 2007). Thus, we do

[29] Differential attrition occurs if attrition systematically differs across the treatment and control groups. Selective attrition occurs when the mean of baseline test scores differs, conditional on treatment status (see Ghanem et al. 2019).

not plan to apply a separate p-value correction (such as Romano and Wolf (2005) or Westfall and Young (1993); see List et al. (2019)).

### 4.3 Cost-effectiveness analysis

The final cost is calculated as the total cost, which is calculated using the present value streams of the cost data listed in Section 3.4, divided by the total impact over the life of the program.

The total impact over the life of the program is calculated as the total impact for the entire group, i.e., the difference between the average test scores in the treatment and comparison groups over the (approximately 1.5-year) life of the program, noting that the estimate for the comparison group is that which would occur had the treatment areas not experienced the program.

## 5 Baseline Results

### 5.1 Statistical power

Using the baseline information, in Figure 1, we show the study's minimal detectable effect (MDE) size for primary outcomes, against varying levels of attrition we may observe later in the study (once we try to follow up on the same sample of students). Even with high levels of attrition (30%), the study would nevertheless continue to be very well powered. In this scenario, we would be able to detect even relatively small effects on children's math scores (of less than 0.15 standard deviations).

More technically, our calculations consider a power of 0.8, given the following assumptions: We use the baseline sample of 4,026 children, from the study's 98 GPs and 292 schools. Based on recent work in Rajasthan Government schools (Ganimian et al. 2017), we expect that baseline data will explain 60% of the outcome variables' variance, at endline—however, we also show more conservative scenarios (for a range from 60% to 10%). Our calculations moreover account for the clustering of standard errors; in doing so, we use the baseline results to calculate the intra-cluster (i.e., intra-GP) correlation (ICC).[30]

[Insert Figure 1 here.]

[30] We calculate the ICC for residuals from a regression of math scores on students' age, gender, and randomization strata fixed effects. The resulting ICC is 0.0359. Note that this approach is conservative as our final analyses will moreover include baseline test scores, as an additional covariate.

### 5.2 Balance of student and school characteristics, at baseline

In this sub-section, we investigate whether the study's randomization strategy created groups whose observable baseline characteristics are balanced.

We begin by visually inspecting the baseline distribution of test scores, by treatment group. Figure 2 presents balance estimates for a combined, overall math score, from the written and oral tests (standardized, scaled with a 2PL IRT model).[31] As expected, we do not see any notable differences across the two groups of students, in terms of their baseline performance distributions.

[Insert Figure 2 here.]

We continue by providing formal balance tests, at the student (Table 2) and school (Table 3) levels. Both tables provide additional support for our finding that the treatment and control groups are balanced.

[Insert Table 2 here.]

[Insert Table 3 here.]

In summary, we interpret these results as confirmation that the randomization "worked"; we are confident that the random assignment created comparable groups of students.

## 6 Timeline

The duration of the study is 1.5 years. Baseline data were collected in November of 2018, after which GKA program implementation launched in the treatment schools, in December of 2018. Between January and February of 2019, J-PAL South Asia conducted one round of classroom observations in treatment and control schools, along with process monitoring in treatment schools.[32] Midline assessment data have been collected in the following school year, in September 2019.[33] During this school year, J-PAL South Asia will also conduct three rounds of classroom observations in treatment and control schools. Endline assessment data will be collected in February of 2020, which will be used to estimate the impact of the program after approximately 1.5 years of operation. A detailed timeline is provided in Table 4.

[Insert Table 4 here.]

[31] For this analysis, we combine all items of the standardized math assessment (administered orally and as a written test), and an additional, binary ASER "item", which indicates whether a student performed at the subtraction or division levels. We provide a more detailed breakdown of student performance, including formal tests, just below.

[32] The study's team of Principal Investigators have not yet accessed these data.

[33] Data entry is not yet complete.

# 7 References


Abadie, A., Athey, S., Imbens, G., Wooldridge, J., 2017. When Should You Adjust Standard Errors for Clustering? Technical Report w24003. National Bureau of Economic Research. Cambridge, MA. doi:10.3386/w24003.

Aiyar, Y., Bhattacharya, S., 2016. The Post Office Paradox: A Case Study of the Block Level Education Bureaucracy. Economic & Political Weekly 51, 61–69. URL: http://pubdocs.worldbank.org/en/734301481774805564/The-Post-Office-Paradox-A-Case-Study-of-the-Block-Level-Education-Bureaucracy.pdf.

Allcott, H., 2015. Site Selection Bias in Program Evaluation. The Quarterly Journal of Economics 130, 1117–1165. doi:10.1093/qje/qjv015.

Anderson, M.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, 1481–1495. doi:10.1198/016214508000000841.

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

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

Athey, S., Imbens, G., 2017. The Econometrics of Randomized Experiments, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 73–140. doi:10.1016/bs.hefe.2016.10.003.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M., 2017a. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives 31, 73–102. doi:10.1257/jep.31.4.73.

Banerjee, A., Chassang, S., Snowberg, E., 2017b. Decision Theoretic Approaches to Experiment Design and External Validity, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 73–140. doi:10.1016/bs.hefe.2016.10.002.

Barrera-Osorio, F., García, S., Rodríguez, C., Sánchez, F., Arbeláez, M., 2018. Concentrating Efforts on Low-Performing Schools: Impact Estimates from a Quasi-Experimental Design. Economics of Education Review 66, 73–91. doi:10.1016/j.econedurev.2018.07.001.

Barro, R.J., 1991. Economic Growth in a Cross Section of Countries. The Quarterly Journal of Economics 106, 407–443. doi:10.2307/2937943.

de Barros, A., 2018. Do Students Benefit from Blended Instruction? Experimental Evidence from India. URL: https://www.socialscienceregistry.org/trials/3665/history/45255.

Birnbaum, A., 1968. Some Latent Trait Models and Their Use in Inferring an Examinee's Ability, in: Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 397–479.

Brown, A.J., Goin, L., Gregory, N., Hoffmann, K., Smith, K., 2015. Evaluating Financial Products and Services in the US: A Toolkit for Running Randomized Controlled Trials. URL: https://www.poverty-action.org/sites/default/files/publications/Evaluating%20Financial%20Products%20and%20Services%20in%20the%20US%20-%20A%20Toolkit%20for%20Running%20RCTs.pdf.

Bruhn, M., McKenzie, D., 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. American Economic Journal: Applied Economics 1, 200–232. doi:10.1257/app.1.4.200.

Carril, A., 2017. Dealing with Misfits in Random Treatment Assignment. The Stata Journal 17, 652–667. doi:10.1177/1536867X1701700307.

Cilliers, J., Fleisch, B., Prinsloo, C., Taylor, S., 2019. How to Improve Teaching Practice? Experimental Comparison of Centralized Training and In-classroom Coaching. Journal of Human Resources (published before print), 0618–9538R1. doi:10.3368/jhr.55.3.0618-9538R1.

Conn, K.M., 2017. Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. Review of Educational Research 87, 863–898. doi:10.3102/0034654317712025.

Dhaliwal, I., Duflo, E., Glennerster, R., Tulloch, C., 2013. Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries - A General Framework with Applications for Education, in: Glewwe, P. (Ed.), Education Policy in Developing Countries. The University of Chicago Press, Chicago, pp. 285–338.

Dhar, D., Jain, T., Jayachandran, S., 2018. Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India. Working Paper 25331. National Bureau of Economic Research. doi:10.3386/w25331.

Escueta, M., Quan, V., Nickow, A.J., Oreopoulos, P., 2017. Education Technology: An Evidence-Based Review. Working Paper 23744. National Bureau of Economic Research. doi:10.3386/w23744.

Evans, D.K., Popova, A., 2016. What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. The World Bank Research Observer 31, 242–270. doi:10.1093/wbro/lkw004.

Friedman, W., Gerard, F., Ralaingita, W., 2010. International independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) program in Mali, mid-term report. Research Triangle Park, NC: RTI International .

Ganimian, A.J., de Barros, A., Muralidharan, K., 2017. Do Students Benefit from Personalized Learning? Experimental Evidence from India. URL: https://www.socialscienceregistry.org/trials/2459.

Ganimian, A.J., Murnane, R.J., 2016. Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. Review of Educational Research 86, 719–755. doi:10.3102/0034654315627499.

Ghanem, D., Hirshleifer, S., Ortiz-Becerra, K., 2019. Testing Attrition Bias in Field Experiments. URL: https://www.dropbox.com/s/5em8isbzghl46qh/GhanemHirshleiferOrtiz_Draft.pdf?dl=0.

Glennerster, R., 2017. The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 175–243. doi:10.1016/bs.hefe.2016.10.002.

Goodnight, M.R., Bobde, S., 2018. Missing children in educational research: investigating school-based versus household-based assessments in India. Comparative Education 54, 225–249. doi:10.1080/03050068.2017.1383085.

Gugerty, M.K., Karlan, D., 2018. The Goldilocks Challenge: Right-Fit Evidence for the Social Sector. Oxford University Press, Oxford.

Hanushek, E.A., Woessmann, L., 2008. The Role of Cognitive Skills in Economic Development. Journal of Economic Literature 46, 607–668. doi:10.1257/jel.46.3.607.

Heß, S., 2017. Randomization Inference with Stata: A Guide and Software. The Stata Journal: Promoting communications on statistics and Stata 17, 630–651. doi:10.1177/1536867X1701700306.

J-PAL, 2017. J-PAL Research Protocols. URL: https://drive.google.com/file/d/0B97AuBEZpZ9zZDZZbV9abllqSFk/view.

Jacob, B., Rothstein, J., 2016. The Measurement of Student Ability in Modern Assessment Systems. Journal of Economic Perspectives 30, 85–108. doi:10.1257/jep.30.3.85.

Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental Analysis of Neighborhood Effects. Econometrica 75, 83–119. doi:10.1111/j.1468-0262.2007.00733.x.

Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking. 3rd ed., Springer, New York, NY.

Krueger, A.B., Lindahl, M., 2001. Education for Growth: Why and For Whom? Journal of Economic Literature 39, 1101–1136. doi:10.1257/jel.39.4.1101.

Lai, F., Luo, R., Zhang, L., Huang, X., Rozelle, S., 2015. Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. Economics of Education Review 47, 34–48. doi:10.1016/j.econedurev.2015.03.005.

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M., Rozelle, S., 2013. Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi. Journal of Development Effectiveness 5, 208–231. doi:10.1080/19439342.2013.807862.

Lee, D.S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies 76, 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x.

Lin, W., Green, D.P., 2016. Standard Operating Procedures: A Safety Net for Pre-Analysis Plans. PS: Political Science & Politics 49, 495–500. doi:10.1017/S1049096516000810.

Lin, W., Green, D.P., Coppock, A., 2016. Standard Operating Procedures for Don Green's Lab at Columbia. URL: https://github.com/acoppock/Green-Lab-SOP.

List, J.A., Shaikh, A.M., Xu, Y., 2019. Multiple hypothesis testing in experimental economics. Experimental Economics (published before print), 1–21. doi:10.1007/s10683-018-09597-5.

Lucas, A.M., McEwan, P.J., Ngware, M., Oketch, M., 2014. Improving Early-Grade Literacy in East Africa: Experimental Evidence From Kenya and Uganda. Journal of Policy Analysis and Management 33, 950–976. doi:10.1002/pam.21782.

Majerowicz, S., Montero, R., 2018. Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru. URL: https://scholar.harvard.edu/files/smajerowicz/files/coaching.pdf.

Mankiw, N.G., Romer, D., Weil, D.N., 1992. A Contribution to the Empirics of Economic Growth. The Quarterly Journal of Economics 107, 407–437. doi:10.2307/2118477.

Sala-i Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. American Economic Review 94, 813–835. doi:10.1257/0002828042002570.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., Rajani, R., 2019. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. The Quarterly Journal of Economics 134, 1627–1673. URL: http://academic.oup.com/qje/article/134/3/1627/5479257, doi:10.1093/qje/qjz010.

Molina, E., Fatima, S.F., Ho, A., Hurtado, C.M., Wilichowksi, T., Pushparatnam, A., 2018. Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool. Working Paper 8653. The World Bank. Washington, D.C. doi:10.1596/1813-9450-8653.

Muralidharan, K., Singh, A., Ganimian, A.J., 2019. Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. American Economic Review 109, 1426–1460. doi:10.1257/aer.20171112.

Popova, A., Arancibia, V., Evans, D.K., 2016. Training Teachers on the Job: What Works and How to Measure it. Working Paper 7834. The World Bank. Washington, D.C. URL: https://www.openknowledge.worldbank.org/handle/10986/25150.

Romano, J.P., Wolf, M., 2005. Stepwise Multiple Testing as Formalized Data Snooping. Econometrica 73, 1237–1282. doi:10.1111/j.1468-0262.2005.00615.x.

Sabarwal, S., Evans, D.K., Marshal, A., 2014. The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone. Policy Research Working Paper 7021. The World Bank. URL: http://documents.worldbank.org/curated/en/806291468299200683/pdf/WPS7021.pdf.

Samejima, F., 1973. A Comment on Birnbaum's Three-Parameter Logistic Model in the Latent Trait Theory. Psychometrika 38, 221–233. doi:10.1007/BF02291115.

Stocking, M.L., Lord, F.M., 1983. Developing a Common Metric in Item Response Theory. Applied Psychological Measurement 7, 201–210. doi:10.1177/014662168300700208.

UNESCO Institute for Statistics, 2018. Data for the Sustainable Development Goals. URL: http://uis.unesco.org/.

Vaijayanti, K., Mondal, A., 2015. SDMCs in Karnataka: Analysing the quality of SDMC meetings in Hoskote, Kushtagi and Mundargi. URL: https://akshara.org.in/wp-content/uploads/Research-and-Evaluation-Summary-Report-of-SDMCs-2nd-Dec15.pdf.

Vivalt, E., 2019. How Much Can We Generalize From Impact Evaluations? Unpublished manuscript. Australian National University. Canberra, Australia. URL: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf.

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

Wolf, S., Aber, J.L., Behrman, J.R., Tsinigo, E., 2019. Experimental Impacts of the "Quality Preschool for Ghana" Interventions on Teacher Professional Well-being, Classroom Quality, and Children's School Readiness. Journal of Research on Educational Effectiveness 12, 10–37. doi:10.1080/19345747.2018.1517199.

Young, A., 2019. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. The Quarterly Journal of Economics 134, 557–598. doi:10.1093/qje/qjy029.

# 8 Appendices

## 8.1 Appendix A: Baseline field operations and quality control

### 8.1.1 Field operations

The study's baseline was successfully conducted between 14 November 2018 and 29 November 2018. The tests were administered by staff hired and trained by J-PAL South Asia. The field team consisted of 48 enumerators, 12 supervisors, two logistics support staff, and two field monitors, allocated equally across the two districts. The team was provided classroom and field training on the data collection instruments and field protocols, for seven days. Thereafter, a "false start" of the survey was conducted in 24 non-sampled schools prior to the commencement of baseline data collection (cf. Brown et al. 2015). A team of two enumerators typically visited one school per day. However, in blocks with lower rates of enrolment, a single enumerator was sufficient to cover a school. The average productivity across both districts was 24 schools per day.

### 8.1.2 Quality control

The following types of data quality checks were conducted in adherence with J-PAL's data quality standards (J-PAL 2017).

**Checks programmed into the survey form.** The tablet-based test was conducted using SurveyCTO—a data collection software that allows for offline operations. Where relevant, logical checks were embedded into the survey form to prevent manual errors. For example, enumerators were asked to enter every student's unique ID twice, and they were prevented from moving forward if there was a mismatch between the two IDs. Similarly, the form was programmed to disallow enumerators from entering the same STS ID for two different students in the same form.

**Spot-checks and accompaniments.** Supervisors were allocated to teams of four enumerators, and district-level monitors were primarily responsible for making announced (accompaniments) and unannounced (spot-checks) visits to observe data collection. A separate monitoring survey was developed to track and report violations in protocols. For example, supervisors and monitors were asked to mark whether the enumerator was clear in providing instructions to students or whether the enumerator left the test papers unattended.

**Weekly monitoring and de-briefs.** Any reported violations in protocols were discussed with the field team on a weekly basis. The tablet-based data were also tracked weekly for enumerator productivity, attendance, enrolment, missing student identifiers, and consent refusals.

## 8.2 Appendix B: Test design and validity evidence

We measure student achievement in mathematics with a test that seeks to capture what students know and can do in this area, with direct reference to their schools' official Kannada-medium curriculum. The assessment is summative and of low stakes, both for test takers and for the study's schools. These tests are administered under the supervision of the research team at baseline, midline, and endline. In this appendix, we present validity evidence for the test's contents and for the test's internal coherence.

### 8.2.1 Content validity

The tests are administered on paper, as multiple-choice tests, and contain 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 improve the test's content validity through four strategies, as follows. First, prior to the baseline, we discussed the test blueprint and content with the implementing organization.[34] Secondly, for each round of assessments, we engage with an external panel of subject matter experts.[35] Third, we map each test question to the official schoolbooks used in Karnataka. Fourth, we accompany each round of test development with field pilots (out-of-sample), to further assess the local relevance of questions and their use of Kannada language.

### 8.2.2 Internal coherence and reliability

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

[34] To make sure the test administration remains impartial and unbiased, we do not repeat this strategy for the midline and endline tests.

[35] The panel consists of former teachers and curriculum experts. The panel does not include staff of the implementing organization.

of the performance distribution.[36] Taken together, we therefore do not expect that floor or ceiling effects limit the test's validity, overall.

Next, we turn to the *range* of ability covered by test questions. Table B1 displays the $\alpha$ and $b$ parameters for the 32 test questions, as per a 2PL IRT model.[37] 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.[38] From this analysis, we include 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 are as a group. The Cronbach's alpha (C$\alpha$) is a widely used measure of reliability in psychometric testing. The C$\alpha$ is a function of the number of items in a test, the covariance between pairs of items, and the variance of the total score. The theoretical value of C$\alpha$ varies from 0 to 1, with a rule of thumb of 0.7 or higher suggesting that the test is reliable. In this study, the C$\alpha$ is 0.91 for the 32 written items. We thus conclude that our instrument is highly reliable overall.

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

[36] This finding corresponds to our previous analysis of Figure 2, whose kernel density plot of tests scores did not indicate floor or ceiling effects.

[37] A 3PL model did not converge for the baseline.

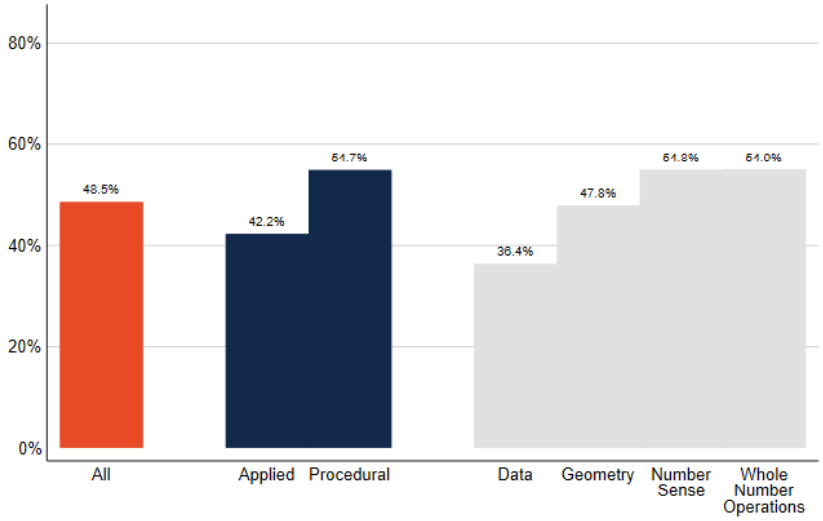
[38] We will keep this item with low discrimination (Q1140) in the baseline assessment. However, we will not repeat the item in our midline or endline assessments (i.e., it will not serve as an "anchor item").

**Table B1:** *Item characteristics*

| Item | □ (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 |

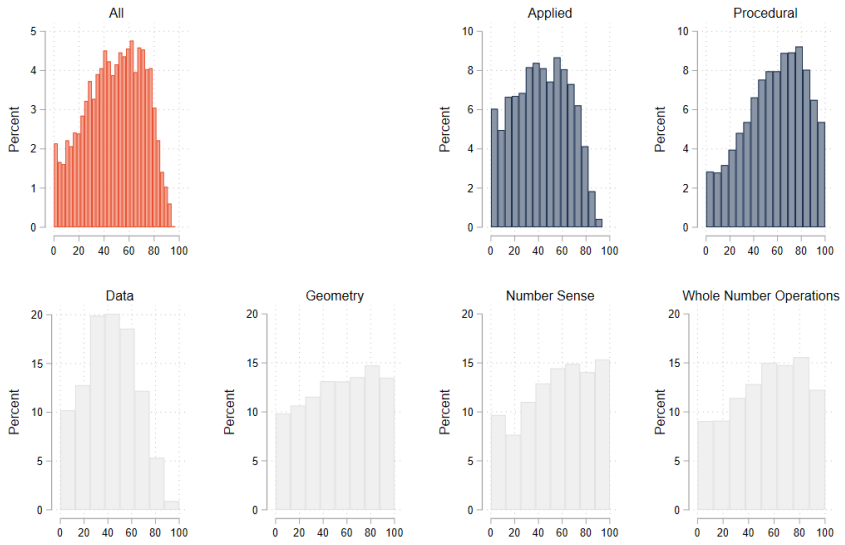
Notes: This table reports on items' discrimination and difficulty parameters, for the baseline as per a 2PL IRT model. Item numbers (in parentheses) refer to study-internal question IDs. Items are sorted by difficulty; items cover a wide range of difficulties. With the exception of one item (Q1140),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

**Figure B1:** *Mean percentage of items solved correctly* (*Baseline, Nov. 2018*)



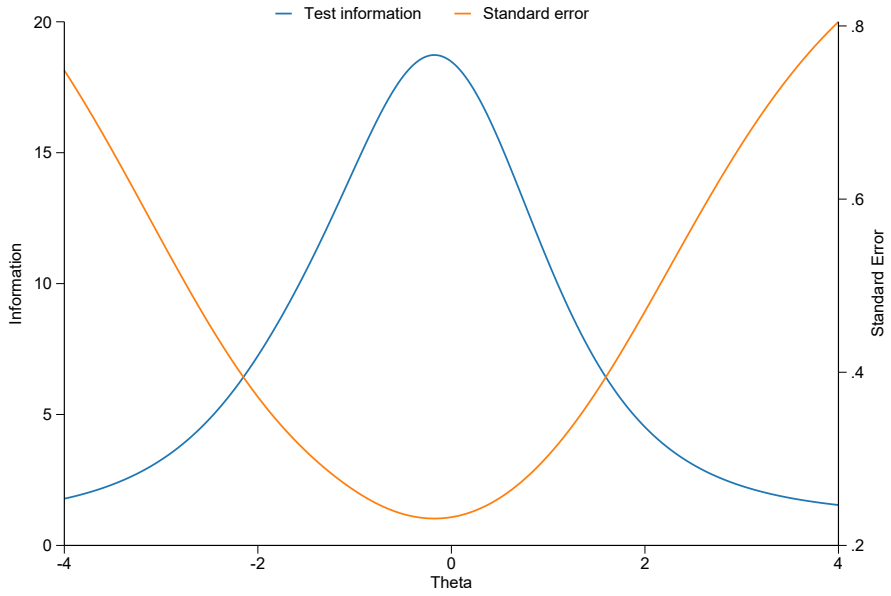

Notes: This figure provides the mean percentage of test questions students solved correctly during baseline (overall, by cognitive domains, and by content domains).

**Figure B2:** *Distribution of percentage of items solved correctly* (*Baseline, Nov. 2018*)




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

**Figure B3:** *Test information function (TIF)*




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

# 9 Administrative Information

## 9.1 Funding

We gratefully acknowledge the funding provided by the Omidyar Network for this project. This research did not receive any additional grants from funding agencies in the public, commercial, or not-for-profit sectors.

## 9.2 Institutional Review Board approval

All necessary ethics approvals are in place. The study has been approved by the University of Minnesota Human Research Protection Program (Study 4101). In India, the study has been approved by the Institute for Financial Management and Research (IFMR) Human Subjects Committee.

## 9.3 Declaration of interest

None.

## 9.4 Acknowledgments

The authors thank J-PAL SA and its team of field staff. Sandhya Seetharaman, Prajwal Shenoy, and Anuja Venkatachalam have provided excellent research assistance (Shenoy) and outstanding research management (Seetharaman and Vankatachalam). 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. de Barros thanks Alejandro Ganimian for helpful comments. The usual disclaimers apply.

# 10 Tables and Figures

**Table 1:** *Measures of implementation fidelity, program outputs, and take-up*

| Description of Measures | Source |
|---|---|
| **Teaching inputs for activity-based instruction, and related training** | |
| BEO's knowledge of the GKA program | BEO survey |
| Negative feedback received about the GKA program | BEO survey |
| Positive feedback received about the GKA program | BEO survey |
| Whether teachers used GKA TLMs | Classroom observations |
| Which GKA TLMs teachers used | Classroom observations |
| Whether group activities were conducted | Classroom observations |
| If the school receives GKA, whether any teachers attended training for GKA | School survey |
| Whether the teacher who received GKA training teaches Math or another subject | School survey |
| If Grade 4 Math teachers know about GKA | Teacher survey |
| If Grade 4 Math teachers received training for GKA | Teacher survey |
| Number of days of training | Teacher survey |
| If Grade 4 Math teachers received the GKA kit | Teacher survey |
| If Grade 4 Math teachers received training on how to use the GKA kit | Teacher survey |
| How regularly teacher uses the GKA kit for Grade 4 Math | Teacher survey |
| If teachers use the GKA kit for other Grades | Teacher survey |
| If other teachers received GKA training if the Math teacher did not receive training | Teacher survey |
| How often teachers conduct group activities for Grade 4 | Teacher survey |
| Challenges in implementing GKA | Teacher survey |
| Last time the teacher's classroom was visited by Akshara | Teacher survey |
| Teacher perception of impact of GKA | Teacher survey |
| **Community events ("GP contests")** | |
| Usefulness of the GP contest in addressing math performance | BEO survey |
| Steps taken by the community/school as result of GP contest | BEO survey |
| Study school participation in the GP contest | GP contest survey |
| Students from sampled schools participating in the GP contest | GP contest survey |
| Parents of students were present at the GP contest | GP contest survey |
| Parents' knowledge of the GKA program | Parent survey |
| If the school has participated in any GP contests | School survey |
| Student participation in GP contest | Student survey |
| If teachers have heard of GP contests | Teacher survey |
| Whether report card was received after GP contests | Teacher survey |
| Usefulness of the GP contest in addressing math performance | Teacher survey |
| Teachers' report of how involved the parents of their students are involved in GP contests | Teacher survey |
| Steps taken by the community/school as result of GP contest | Teacher survey |
| Usefulness of the GP contest in addressing math performance | HM survey |
| Steps taken by the community/school as result of GP contest | HM survey |

Notes: The parent survey is for a random sub-sample of 4 students per school. The student survey is for a random sub-sample of 8 students per school. BEO stands for "Block Education Officer". GKA stands for "Ganitha Kalika Andolana" (the program evaluated in this study). GP stands for "Gram Panchayat". TLM stands for "Teaching and Learning Materials".

**Table 2:** *Balance tests, student level*

| | Control | | Treatment | | Difference |
|---|---|---|---|---|---|
| Variable | N/[Clusters] | Mean/[SD] | N/[Clusters] | Mean/[SD] | Coef./(s.e.) |
| | (1) | (2) | (3) | (4) | (5) |
| Math Score (2PL, std.) | 1948 | 0.031 | 2078 | -0.029 | -0.043 |
| | [49] | [3.355] | [49] | [2.448] | (0.051) |
| Student ASER level | 1898 | 2.295 | 2039 | 2.297 | -0.004 |
| | [49] | [1.397] | [49] | [1.849] | (0.032) |
| Percentage correct (all items) | 1948 | 0.532 | 2078 | 0.518 | -0.010 |
| | [49] | [0.743] | [49] | [0.535] | (0.011) |
| Percentage correct (oral test, includes ASER as 1 'item') | 1898 | 0.693 | 2039 | 0.694 | -0.000 |
| | [49] | [0.460] | [49] | [0.500] | (0.003) |
| Percentage correct (written test) | 1948 | 0.493 | 2078 | 0.476 | -0.012 |
| | [49] | [0.827] | [49] | [0.581] | (0.014) |
| Percentage correct (applied, written) | 1948 | 0.432 | 2078 | 0.414 | -0.013 |
| | [49] | [0.846] | [49] | [0.612] | (0.015) |
| Percentage correct (procedural, written) | 1948 | 0.556 | 2078 | 0.539 | -0.012 |
| | [49] | [0.823] | [49] | [0.564] | (0.013) |
| Percentage correct (whole number operations, written) | 1948 | 0.551 | 2078 | 0.546 | -0.002 |
| | [49] | [0.887] | [49] | [0.665] | (0.015) |
| Percentage correct (numbers, written) | 1948 | 0.558 | 2078 | 0.538 | -0.014 |
| | [49] | [0.867] | [49] | [0.585] | (0.016) |
| Percentage correct (data, written) | 1948 | 0.376 | 2078 | 0.353 | -0.019 |
| | [49] | [0.706] | [49] | [0.530] | (0.013) |
| Percentage correct (geometry, written) | 1948 | 0.487 | 2078 | 0.469 | -0.012 |
| | [49] | [0.954] | [49] | [0.731] | (0.017) |
| Gender: Female | 1898 | 0.532 | 2039 | 0.529 | -0.002 |
| | [49] | [0.538] | [49] | [0.694] | (0.017) |
| Student age (as of 31-Dec-18) | 1901 | 9.140 | 2044 | 9.154 | 0.012 |
| | [49] | [1.363] | [49] | [1.471] | (0.024) |
| F-test of joint significance (F-stat) | | | | | 1.047 |
| F-test, number of observations | | | | | 3857 |

Notes: The value displayed for *t*-tests are the differences in the means across the groups.
The value displayed for *F*-tests are the *F*-statistics.
Standard deviations (SD) and number of clusters in brackets. Standard errors (s.e.) in parentheses, clustered at the Gram Panchayat level.
Randomization strata fixed effects are included in all estimation regressions.
***, ** and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 3:** *Balance tests, school level*

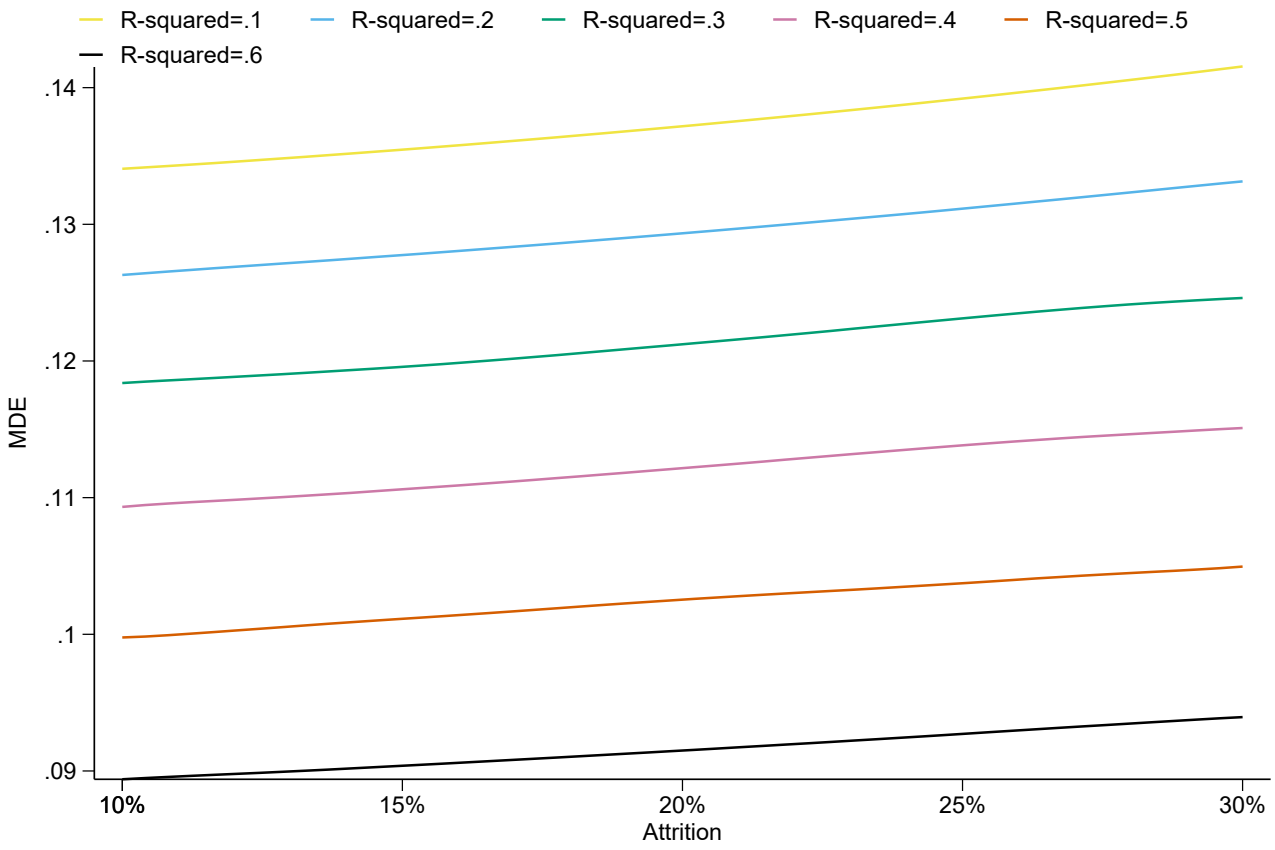
| | Control | | Treatment | | Difference |
|---|---|---|---|---|---|
| Variable | N/[Clusters] | Mean/[SD] | N/[Clusters] | Mean/[SD] | Coef./(s.e.) |
| | (1) | (2) | (3) | (4) | (5) |
| Percent of students appeared and passed primary exam | 133 | 0.317 | 132 | 0.328 | 0.084* |
| | [49] | [0.457] | [49] | [0.394] | (0.050) |
| Female students (percentage) | 146 | 0.497 | 146 | 0.506 | -0.010 |
| | [49] | [0.087] | [49] | [0.107] | (0.013) |
| Percentage OBC | 146 | 0.645 | 146 | 0.655 | -0.001 |
| | [49] | [0.247] | [49] | [0.276] | (0.033) |
| Total number of teachers | 146 | 5.192 | 146 | 5.295 | 0.381 |
| | [49] | [3.126] | [49] | [3.011] | (0.312) |
| No of students per teacher | 146 | 28.710 | 146 | 26.659 | -1.650 |
| | [49] | [21.980] | [49] | [16.265] | (1.571) |
| Female teachers (percentage) | 146 | 0.458 | 146 | 0.414 | -0.028 |
| | [49] | [0.401] | [49] | [0.369] | (0.035) |
| School: Years in service | 146 | 73.274 | 145 | 69.669 | -3.158 |
| | [49] | [27.626] | [49] | [21.380] | (3.236) |
| School is co-ed (vs. single-sex) | 146 | 0.856 | 146 | 0.932 | 0.119** |
| | [49] | [0.457] | [49] | [0.312] | (0.051) |
| Percentage of classrooms needing minor repair | 146 | 0.144 | 146 | 0.128 | -0.003 |
| | [49] | [0.188] | [49] | [0.187] | (0.017) |
| Percentage of classrooms needing major repair | 146 | 0.126 | 146 | 0.147 | 0.015 |
| | [49] | [0.219] | [49] | [0.201] | (0.019) |
| No. toilets / students | 146 | 0.027 | 146 | 0.026 | -0.001 |
| | [49] | [0.030] | [49] | [0.025] | (0.001) |
| Boundary wall is inexistent or incomplete | 146 | 0.534 | 146 | 0.630 | 0.198*** |
| | [49] | [0.570] | [49] | [0.545] | (0.067) |
| School has tap water | 146 | 0.589 | 146 | 0.603 | -0.023 |
| | [49] | [0.606] | [49] | [0.661] | (0.058) |
| Computers / no. of students | 146 | 0.014 | 146 | 0.011 | -0.001 |
| | [49] | [0.027] | [49] | [0.024] | (0.002) |
| Received a school maintenance grant | 146 | 0.911 | 146 | 0.884 | -0.049 |
| | [49] | [0.422] | [49] | [0.435] | (0.051) |
| F-test of joint significance (F-stat) | | | | | 1.416 |
| F-test, number of observations | | | | | 264 |

Notes: The values displayed for column (5) are coefficients from regressing each variable on the treatment indicator.
The value displayed for *F*-tests are the *F*-statistics.
Standard deviations (SD) and number of clusters in brackets. Standard errors (s.e.) in parentheses, clustered at the Gram Panchayat level.
***, ** and * indicate significance at the 1, 5, and 10 percent critical level.

**Table 4:** *Timeline of the study*

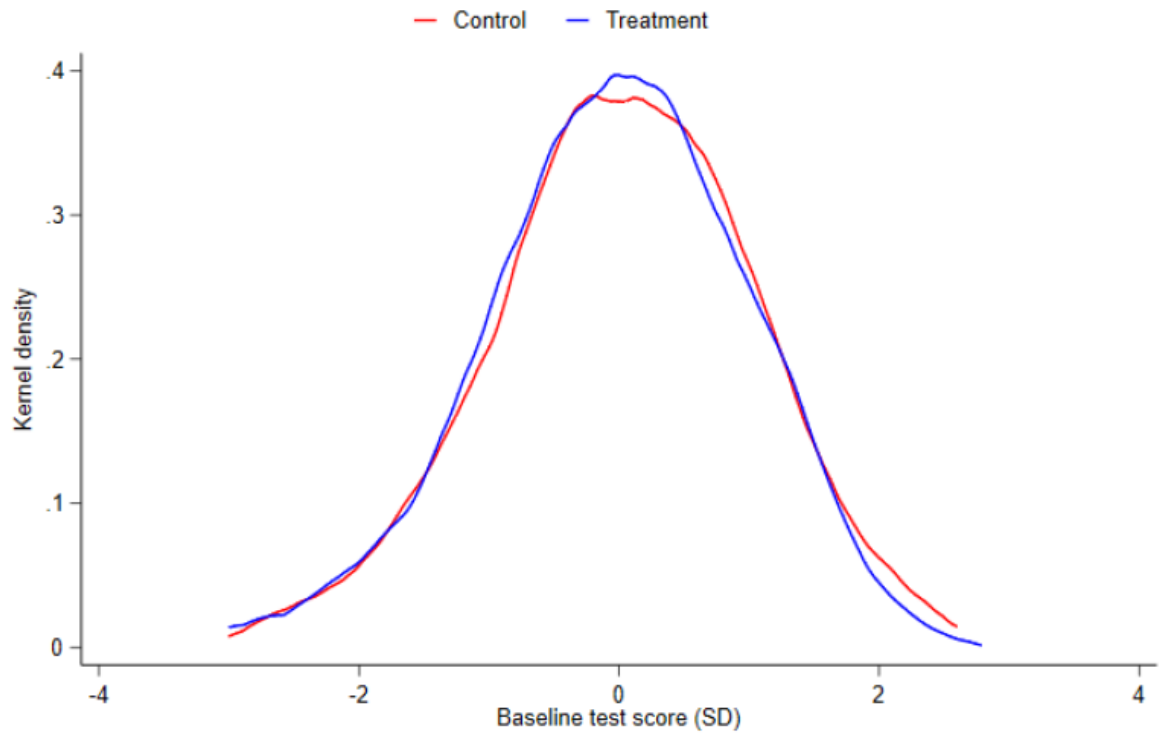
| Start | End | Event |
|---|---|---|
| 1 Aug 2018 | 31 Aug 2018 | ✓ Research team samples Gram Panchayats (GPs) and schools |
| 1 Sep 2018 | 30 Sep 2018 | ✓ Research team secures funding from Omidyar |
| | | ✓ Research team, GoK & Akshara Foundation sign an MoU |
| | | ✓ Research team pilots & finalizes baseline assessments |
| 1 Oct 2018 | 31 Oct 2018 | ✓ Research team secures IRB approval |
| | | ✓ Research team digitizes the one-on-one tests |
| 1 Nov 2018 | 30 Nov 2018 | ✓ Research team administers baseline assessments |
| | | ✓ Research team conducts the randomization of GPs |
| | | ✓ Research team pilots the process monitoring instruments |
| | | ✓ Research team & Akshara Foundation sign an MoU |
| 01 Dec 2018 | 31 Dec 2018 | ✓ Research team digitizes data from the paper-based student assessments |
| | | ✓ Akshara Foundation rolls out GKA in treatment GPs |
| 01 Jan 2019 | 28 Feb 2019 | ✓ Research team conducts classroom observations & process monitoring |
| 01 Jun 2019 | 30 Jun 2019 | ✓ Research team pilots and finalizes the midline assessments |
| 01 Jul 2019 | 31 Aug 2019 | ✓ Research team conducts classroom observations & process monitoring |
| 01 Sep 2019 | 30 Sep 2019 | ✓ Research team administers midline assessments |
| 01 Oct 2019 | 30 Nov 2019 | ✓ Research team conducts classroom observations & process monitoring |
| 01 Dec 2019 | 31 Jan 2020 | ✓ Research team conducts classroom observations & process monitoring |
| | | ✓ Research team pilots and finalizes endline assessments |
| 01 Feb 2020 | 28 Feb 2020 | ✓ Research team administers endline assessments |
| 01 March 2020 | 01 Oct 2020 | • Research team analyzes study results, and disseminates results |

**Figure 1:** *Power calculations*




Notes: This figure provides Minimal Detectable Effects (MDE) by attrition rates, for a range of six different values of $R^2$ (the share of endline variance explained by baseline covariates). 49 control GPs, 49 treatment GPs. 3 schools per GP. Power=0.8. Intra-cluster correlation as per baseline results, within GPs, for residuals from a regression of math scores on students' age, gender, and randomization strata fixed effects.

**Figure 2:** *Balance at baseline, overall test score (Nov. 2018)*




Notes: This figure provides kernel density plots for the overall baseline test score, separately for the control and for the treatment group.