Pre-analysis plan for Mindspark Rajasthan

Karthik Muralidharan and Abhijeet Singh

Abstract

This note outlines the research questions, study design and analysis plans for an evaluation of the Mindspark Computer-aided Learning (CAL) software platform in government schools in Rajasthan. The trial was started in October 2017 and the endline data from 2.5 years of treatment was collected in February-March 2020. This note is being written and uploaded after the collection of all data but before data from the March 2020 endline assessments has become available to the research team.

I. Intervention details

Study Background

Developing countries have made impressive progress in improving school enrolment and completion in the last two decades. Yet, their productivity in converting education investments of time and money into human capital remains very low. This proposal focuses on the increased use of technology in education, which is frequently proposed for improving this low productivity, but where the evidence base for the effectiveness of such policies remains mixed.

This research study builds on promising results reported in Muralidharan, Singh and Ganimian (2019, MSG hereafter) who evaluated a technology-led intervention (Mindspark) providing personalized instruction to students which is individually customized to their level of achievement in after-school settings in Delhi. MSG documented large learning gains, which did not vary significantly by baseline test scores, gender, or household socioeconomic status. The intervention was cost-effective both in terms of money and scheduled time of instruction.

The MSG (2019) evaluation in Delhi, which shall be the reference for this current evaluation, was framed explicitly as a "proof-of-concept" study. It focused on middle school students in Delhi, who applied to be a part of the intervention in an after-school setting, delivered at a small scale and with high fidelity directly by the firm that developed Mindspark (Educational Initiatives India) over half a school-year. The current research project aims to evaluate the effectiveness of such an intervention at a larger scale, directly in government schools, with students in a broader range of grades and for a longer duration. Further, the current study features the use of education technology to *substitute* in-class instruction time as opposed to the setting in MSG that evaluated a *supplemental* intervention.

Intervention

We evaluate an intervention using the Mindspark software (described further below) that provides computer-aided instruction in mathematics and Hindi language for *all* students enrolled in Grades 1-8 in (randomly-assigned) government schools, as part of regularly-scheduled classes in the school time-table.

Mindspark software

Instruction is provided using the Mindspark software which provides personalized instruction to students which is individually customized to their level of achievement. The software aims to address several constraints:

- *Adaptive instruction* addresses the twin constraints of (a) the mismatch between curricular expectations and actual academic levels of students and (b) the wide spread of achievement levels in a single grade, which make delivery of any single curriculum difficult;
- *Engaging interactive content*, often game-based, aims to improve student attention;
- *Real-time assessment and grading* allows for immediate feedback to students and continuous diagnostics about student comprehension
- *Careful analyses of patterns of student errors*, enabled by "big data", allow for precisely targeting content to clarify specific areas of misunderstanding.

Implementation of Mindspark lessons

Mindspark instruction was provided in a dedicated computer lab in each school. Each grade (which is equivalent to a classroom in our setting because no grade had multiple sections) was scheduled to have 1 period per day in the Mindspark Lab (or 5-6 periods per week) with this time being split evenly across math and Hindi. Importantly, these periods came out of regularly scheduled instructional time and thus a Mindspark period for math *replaced* a regularly-scheduled period for math. Teachers were asked to accompany students to the lab and appropriate training was provided to the teachers to carry out day to day functions of the lab. The computer labs are run with the help of lab-in-charges who help students to create their log-in IDs and also help them to log-in on laptops, and help ensure that the computers are always up and running (and interfacing with a tech support team to enable this). In the 2017-18 and 2018-19 academic years, there were dedicated lab-in-charges for every lab. This number was reduced in 2019-20 such that there is typically one lab-in-charge for three schools.

Mindspark classes are scheduled to be subject-specific. In a given class, students are supposed to work on their own on the Mindspark software platform, which customises instruction to their grade level based on an initial diagnostic test and performance over the course of the program. Wherever possible a student worked on their own computer. However, if the number of students became greater than the number of computers, students were paired in groups of 2. Students in a pair were matched based on their diagnostic level. Such pairing was particularly prevalent in middle school grades.

Differentiation from Muralidharan, Singh and Ganimian (2019)

This project is explicitly a follow-up to the MSG(2019) evaluation in Delhi with the shared primary objective of improving student achievement. As described later, analytical choices will also follow closely the procedures used in that study. However, there are many important differences across the studies, which will inform our hypotheses and analyses later. These include:

- **Population treated**: in contrast to MSG(2019), which was focused on an exclusively urban population in Delhi, this study includes both rural and urban populations and in a much more deprived setting than Delhi.

- **Representativeness of population:** The Delhi study was based on a self-selected sample of students who signed up to participate in the after school program (and the study frame comprised less than 10% of students enrolled in the schools in the catchment area). A key difference in current study is that it treats *all the students* enrolled in treated schools.

- **Educational stage:** Whereas MSG(2019) focused on middle school students only, we focus on students from Grades 1-8 and will be able to examine treatment effects at different levels of schooling separately. On a related note, we will be able to examine learning levels and trajectories in the control group over time and quantify treatment effects relative to these.

- **Duration:** Whereas MSG(2019) demonstrated large effects in one-half of a school-year, we can evaluate treatment effects over much longer durations of treatment of up to 2.5 years.

- **Supplementation vs substitution:** The Delhi trial was entirely supplementary to regular school instruction, whereas the current study replaces regular classroom instruction with computer-aided instruction within the school-day. The current study is thus a much better test of whether the Mindspark software increased the productivity of instructional time.

- **Implementing agents:** The Delhi intervention was delivered by staff hired by Educational Initiatives with high fidelity. The current intervention moves the setting to government schools and tests impacts under implementation by regular government teachers and head-teachers.

- **Scale of implementation:** As a proof-of-concept study, MSG(2019) was implemented at a small scale, including 619 students in the study sample, about half of which were assigned to treatment. In contrast, this study is about 20 times the size: average per-school enrolment was about 160 students in Grades 1-8 at baseline in the study schools, thus implying over 12,000 students in the overall study each year.

For all these reasons, the results from the current study are more likely to represent policy-relevant treatment effects based on an at-scale deployment of a personalized education technology intervention (like Mindspark) in a public education system.

II. Study Design

Sampling and Randomization

The study was conducted in Adarsh schools in four districts of Rajasthan – Churu, Dungarpur. Jhunjhunun and Udaipur.[1] 80 schools, split across urban and rural areas in these districts, were selected by the program implementers, Educational Initiatives, as potentially suitable for the rollout of the Mindspark CAL intervention. We randomly allocated 40 of these schools to receive the intervention, with the remaining 40 schools serving as the comparison group, according to the following algorithm:

- Schools in the sampling frame were ranked within-district by their middle school enrolment and assigned into pairs (within district) based on this rank. All districts had an even number of schools in the sampling frame.
- Randomization was stratified at the pair level. One school was selected in each pair (stratum) to be assigned to treatment status with the other serving as a control school.
- The final sample includes 40 treatment and 40 control schools. The final study sample includes 18, 12, 24 and 26 schools in Churu, Dungarpur, Jhunjhunun and Udaipur respectively.

Data

We will use data from four primary sources in our analysis

Independent assessments of student learning

The core objective of the intervention is to raise student learning. We measured this directly using independently-administered tests of student achievement in mathematics and Hindi. These tests were administered by the research team to all students present on the day of testing in Grades 1-8 in the schools, with independent proctoring, in the 80 study schools.

Test papers in each round were designed to be specific to each grade level but designed to capture a broad range of achievement and, in particular, to avoid ceiling and floor effects in measuring student achievement. In middle school grades, all assessments were written, whereas in primary grades, the tests included items assessed orally by interviewers and thus not reliant exclusively of the ability to read and write. Tests were designed to have overlap in items between adjacent grades and across multiple test rounds to allow for linking of assessments on a common metric, both over time and across various grades, using Item Response Theory models.

These independent assessments were conducted four times over the course of the study: a baseline assessment in Oct-Nov 2017 and an end-of-year assessment in Feb-March of each of the three academic years 2017-18, 2018-19 and 2019-20. In the final assessment round in Feb-March 2020, in addition to administering tests to all students in schools, we also followed up and tested any students

[1] Adarsh schools are integrated schools that include all grades from primary to secondary classes. There are 10,712 such schools in Rajasthan.

who were absent on the day of testing in their households. At the time of writing this analysis plan, data from the endline assessment in 2019-20 which shall present our main treatment effect results, is not yet available to the research team.[2]

Data from the Mindspark Computer-Aided-Learning (CAL) system

The Mindspark software collects, in the treatment schools only, detailed individual level data on (a) the assessed grade level of the child based on the diagnostic test and (b) the full record of each specific question attempted by a student and the time spent on the software. Where students are sitting in a pair, they are meant to both log in, following which that particular "session" would be associated with both students. Questions are linked to a number of characteristics such as the skill covered and the grade level of the question.

These data will allow us to compute individual-specific usage and to look at the precise content that each individuals were exposed to, for which we will use similar procedures as in the MSG(2019) evaluation in Delhi.

School-level administrative data

These data will be assembled from all study schools and will cover three main areas: (a) student achievement as measured in school exams for Grades 5 and 8, which are assessed on a common test paper across schools, (b) student attendance data, collected in monthly aggregates and (c) details on school staffing and infrastructure collected in the U-DISE data which covers all recognized schools in the country.

Primary data collection on intermediate processes and behavioural responses

In addition to the above, we conducted extensive primary data collection in 2019-20 to characterize the intervention delivery once the program had stabilized. This was done using the following sets of data collection:

- In-person observations based on structured data collection protocols with snapshots of time use and functioning in Mindspark lab classes (Treatment schools only)
- In-person snapshot-based observations of classroom-teaching in both treatment and control schools
- Teacher interviews, with a particular focus on changes in time use, syllabus completion, pedagogical strategies as a result of Mindspark.
- Student interviews, with a particular focus on engagement in school and, in treatment schools only, on their experience of Mindspark classes.

III. Analysis

Our analytical choices will follow MSG(2019) closely as the "benchmark case".

Primary outcomes

[2] The endline assessments were successfully conducted before the COVID-19 induced shutdown of schools, but data entry was severely delayed because of the lockdown.

Our primary outcome of interest is student test scores, as measured by independent school tests in Feb-March 2020, after 2.5 years of the program being implemented in treatment schools, in math and Hindi language. At the time of writing this analysis plan, these data are not yet available to us (although we have seen and presented test score data from earlier years).

We shall generate these test scores using Item Response Theory models. These will be estimated after pooling item-level data from all rounds of testing and which link across grades and over time, using "anchor items" which were administered in common across test booklets. These choices are similar to MSG (2019).

We shall report both unadjusted and adjusted differences in student achievement between treatment and control groups, always including stratum fixed effects and clustering standard errors at the school level (which is the unit of randomization). These are all Intent-to-Treat estimates.

Our core specification for reporting unadjusted differences in student achievement is

$$Y_{igspt}^{j} = \theta_p + \beta_1 . Treatment_s + \epsilon_{igspt}$$

Where Y_{igspt}^{j} is the test score in subject j for student i in grade g in school s in randomization stratum (pair) p at time t; Treatment is an indicator variable for being in a program school; θ_p is a vector of stratum (school-pair) fixed effects. We will standardize the IRT-test scores within each subject-grade to report treatment effects that are comparable with the education literature. We will also present effects based on pooled IRT estimates. This will also allow us to characterize the treatment effects relative to value-added in the control group (which is not possible when test scores are standardized within grade since the control mean is zero by construction).

We will further report differences in test scores after conditioning for baseline achievement (from Oct 2017). We will present these using individual-level test scores where available; for students who were not present on the day of the baseline test, we will replace the individual-level score with the average for the relevant grade in the school in the baseline.[3] Our specification is presented below:

$$Y_{igspt}^{j} = \theta_p + \beta_1 . Treatment_s + \beta_2 . + Y_{igsp0}^{j} + \ \epsilon_{igspt}$$

We will report these specifications for the end-of-year assessments conducted in Feb-March in each of the three academic years that the program was in place, corresponding to ~0.5, 1.5 and 2.5 years after the program was introduced. To check for sensitivity of results to choices in adjusting for baseline tests, we will also report specifications that control for baseline achievement using (a) classroom average from the baseline alone and (b) classroom averages and individual-level scores.

[3] This will be done so as to not lose individuals who may have been absent on the day of the baseline test but for whom later outcome data is available. For cohorts which entered school in subsequent years, i.e. did not have a score at baseline, we will replace the lagged test score with the average for the preceding cohort.

We will report results for two groups: (a) all individuals tested in any given end-of-year assessment and (b) only those students who were continuously enrolled in the study schools over the period until that assessment. Note that the restriction implied in (b) would drop all students in Grades 1 and 2 in 2019-20 and a substantial fraction of students in Grades 6 and 7.[4]

Main Sub-group analyses

We will also be investigating heterogeneity in treatment effects across a number of dimensions.

Grade levels: We will particularly be interested, in this setting, in investigating treatment effects separately by the grade levels in which students are enrolled. We will thus investigate treatment effects in three sub-groups: Grades 1-2 (early primary), Grades 3-5 (late-primary) and Grades 6-8 (middle school). Heterogeneity across these sub-groups are likely to be salient because students in Grades 1-2 may be too young to benefit from CAL, and those in 6-8 may benefit even more from the customized content given the larger within-grade heterogeneity in learning levels that is likely in higher grades.

Baseline achievement: The personalization of instruction makes it particularly interesting, as in the Delhi trial, to examine heterogeneity in treatment effects by initial position in the baseline distribution of test scores.

We will follow our procedures from the Delhi trial in these investigations and will first examine heterogeneity in treatment effects using non-parametric local linear regressions and, separately, in a linear specification. We will further, in a linear specification, allow the treatment effect to vary by within-grade quantiles of achievement at baseline.

Other covariates: We will also investigate heterogeneity in treatment effects also by gender and SES using linear specifications, interacting the treatment indicator with the relevant covariate, as in MSG (2019).

Supplementary analyses

Heterogeneity within grade and distance from curriculum

In MSG(2019), we used diagnostic test data from the Mindspark system at baseline to document that (a) students were substantially behind appropriate levels of achievement in middle schools, (b) the range within each grade spanned 3-5 grade levels and (c) the deficit between the level of achievement expected by the syllabus and students' actual achievement level seemed to grow across grades. We will present a similar characterization of the sample also in this evaluation, following the template of MSG(2019).

[4] While Adarsh schools have primary school grades as well, a substantial fraction of the students they serve in middle school grades (Grade 6-8) would have transferred from a different primary school after completing Grade 5.

School level tests

We will also investigate program effects on school-based official tests. As shown in MSG(2019), substantial program effects may be missed in tests which only include syllabus-aligned questions in settings where students are very far from grade-level achievement. In this setting, these are administered using a common question paper across schools only in Grades 5 and 8, which will be the relevant groups for us to assess these effects.

We will focus on documenting effects in the 2018-2019 school exams since final exams for the 2019-20 school year were cancelled due to the coronavirus pandemic and school closures.

Behavioural responses to Mindspark implementation

In the present study, unlike the Delhi trial, the Mindspark instruction substitutes (rather than supplements) regular classroom instruction time. As such, we will evaluate the resulting differences in teacher allocation of classroom time and pedagogical practices (from direct observation and teacher reports) as well as substitution of time away from other productive activities such as revision, remedial instruction or instruction in other subjects.

Characterizing program implementation

We will use data from direct lab observations and from teacher and student interviews, conducted in 2019-20, to characterize the implementation of the program. These are descriptive statistics which will be presented separately for Grades 5 and 8 and for math and Hindi.