NBER WORKING PAPER SERIES

DISRUPTING EDUCATION? EXPERIMENTAL EVIDENCE ON TECHNOLOGY-AIDED INSTRUCTION IN INDIA

Karthik Muralidharan
Abhijeet Singh
Alejandro J. Ganimian

Working Paper 22923
http://www.nber.org/papers/w22923

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2016

We thank Abhijit Banerjee, Peter Bergman, Gordon Dahl, Heather Hill, Priya Mukherjee, Chris Walters and several seminar participants for comments. We thank the staff at Educational Initiatives (EI)—especially, Pranav Kothari, Smita Bardhan, Anurima Chatterjee, and Prasid Sreeprakash—for their support of the evaluation. We also thank Maya Escueta, Smit Gade, Riddhima Mishra, and Rama Murthy Sripada for excellent research assistance and field support. Finally, we thank J-PAL's Post-Primary Education initiative for funding this study. The operation of Mindspark centres by EI was funded by the Central Square Foundation, Tech Mahindra Foundation and Porticus. All views expressed are those of the authors and not of any of the institutions they are affiliated with, nor of the National Bureau of Economic Research.

Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India
Karthik Muralidharan, Abhijeet Singh, and Alejandro J. Ganimian
NBER Working Paper No. 22923
December 2016, Revised January 2017
JEL No. C93,I21,O15

ABSTRACT

We present experimental evidence on the impact of a technology-aided after-school instruction program on learning outcomes in middle school grades in urban India, using a lottery that provided students with a voucher to cover program costs. A key feature of the program was its ability to individually customize educational content to match the level and rate of progress of each student. We find that lottery winners had large increases in test scores of 0.36σ in math and 0.22σ in Hindi over just a 4.5-month period. IV estimates suggest that attending the program for 90 days would increase math and Hindi test scores by 0.59σ and 0.36σ respectively. We find similar absolute test score gains for all students, but the relative gain was much greater for academically-weaker students because their rate of learning in the control group was close to zero. We show that the program precisely targets instruction to students' preparation level, thus catering effectively to the very wide variation in student learning levels within a single grade. The program was highly cost-effective, both in terms of productivity per dollar and unit of time. Our results suggest that well-designed technology-aided instruction programs can sharply improve productivity in delivering education.

Karthik Muralidharan
Department of Economics, 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
kamurali@ucsd.edu

Abhijeet Singh
Department of Economics,
University College London
30 Gordon Street, London,
UK
abhijeetsingh1@gmail.com

Alejandro J. Ganimian
Abdul Latif Jameel Poverty Action Lab,
2 Balbir Saxena Marg,
New Delhi, India
aganimian@povertyactionlab.org

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/980

1 Introduction

Developing countries have made impressive progress in improving school enrollment and completion in the last two decades. Yet, their productivity in converting education investments of time and money into human capital remains very low. For instance, in India, over 60% of children aged 6-14 cannot read at the second grade level, despite primary school enrollment rates over 95% (ASER, 2014). Further, there have been very limited improvements in learning outcomes in India in the past decade despite substantial increases in education spending in this period (Muralidharan, 2013). More generally, even in developed countries, productivity growth in the education sector lags the rest of the economy, perhaps because the 'technology' of schooling (classroom-based instruction) has changed very little over time compared to rapid technological progress in other fields (Bosworth, 2005; Pritchett, 2013).

Thus, it is not surprising that increasing the use of technology in instruction is seen as a leading candidate for 'disrupting' the status quo and improving productivity in education (Negroponte et al., 2006; Khan, 2012; Mead, 2016).[1] Yet, the evidence to date appears rather mixed: A recent review of evidence from high-quality studies on the impact of using technology in education globally reports "mixed evidence with a pattern of null results" (Bulman and Fairlie, 2016). Overall, the evidence thus far suggests that realizing the potential of technology-aided education to improve education will require paying careful attention to the *details* of the specific intervention, and the extent to which it alleviates binding constraints to learning.

In this paper, we present experimental evidence on the impact of a technology-led instructional program (called Mindspark) that aimed to leverage technology to improve education by paying sustained attention to such design details. Developed by a leading Indian education firm, the Mindspark program reflects over 10 years of product development; it has been used by over 400,000 students, has a database of over 45,000 test questions, and administers over a million questions across its users every day. A key feature of Mindspark is its ability to use these data to finely benchmark the learning level of every student and dynamically customize the material being delivered to match the level and rate of progress made by each individual student. A second noteworthy feature is its ability to analyze these data to identify patterns of student errors, and precisely target content to alleviate conceptual 'bottlenecks' that may be difficult for teachers to diagnose or address at the individual student level in a classroom setting. Mindspark can be delivered in a variety of settings (in schools, in after-school centers, or through self-guided study); it is platform-agnostic (can be deployed through computers, tablets, or smartphones); and it can be used both online and offline.

[1]A non-exhaustive list of posited channels of impact include using technology to consistently deliver high-quality content that may circumvent limitations in teachers' own knowledge; delivering engaging (often game-based) interactive content that may improve student attention; delivering individually customized content for students; reducing the lag between students attempting a problem and receiving feedback; and, analyzing patterns of student errors to precisely target content to clarify specific areas of misunderstanding.

We evaluate the after-school Mindspark centers in this paper. The centers scheduled six days of instruction per week, for 90 minutes per day; each session was divided into 45 minutes of individual self-driven learning on the Mindspark software and 45 minutes of instructional support from a teaching assistant in groups of 12-15 students.[2] The Mindspark centers aimed to serve students from low-income neighborhoods in Delhi, and charged a modest fee.[3] Our evaluation was carried out in a sample of 619 students recruited for the study from public middle schools in Delhi. Around half of these students were randomly-selected to receive a voucher offering free attendance at Mindspark centers. Students were tested in math and Hindi (language) using independent paper-and-pencil tests before and after the 4.5-month long intervention, with assessments linked using item-response theory (IRT) to be comparable on a common scale across both rounds of testing and across different grades.

We report five main sets of results. First, the dynamic computer-based assessments allow us to provide a more granular description of student learning levels, their distribution within grades, and their evolution over grades than the literature to date. Using these data, we show that student learning levels in our sample are several grade-levels behind their grade-appropriate standard, and that this gap grows by grade. The average grade 6 student is around 2.5 grade levels below sixth grade standards in Math; by grade 9, this deficit increases to 4.5 grade levels. Thus, the default of classroom instruction based on grade-appropriate textbooks is likely to be considerably above the preparation level of most students (especially the lower-achieving ones). Consistent with this, we find that the absolute value-added on our independently-administered tests is close to zero for the bottom-third of students in the control group, and we cannot reject that these students made no academic progress through the school year.

Second, we find that students winning a Mindspark voucher scored 0.36σ higher in math and 0.22σ higher in Hindi relative to students who applied for but did not win the lottery. Relative to the control group, lottery winners experienced twice the test score value-added in math and 2.5 times that in Hindi during the study period of 4.5 months. These are intent-to-treat estimates reflecting an average attendance rate of 58% (including the voucher winners who did not attend regularly). Using the lottery as an instrumental variable, we estimate that attending the Mindspark centers for 90 days (which corresponds to 80% attendance for half a school year), would raise math and Hindi test scores by 0.59σ and 0.36σ respectively.

[2]The teaching assistant focused on helping students with completing homework and with exam preparation, while the instruction was mostly provided by the Mindspark software. Our results may therefore under-estimate the full-potential impact of 90-minutes of "blended" technology-aided learning because the teacher time was not optimized for instruction (see sections 2 and 5 for details).

[3]The online and school-based models require fees that are not affordable for low-income families. The Mindspark centers were set up with philanthropic funding to make the platform more widely accessible, and were located in low-income neighborhoods. However, the funders preferred that a (subsidized) fee be charged, reflecting a widely-held view among donors that cost-sharing is necessary to avoid wasting subsidies on those who will not value or use the product (Cohen and Dupas, 2010). The subsidized fee of Rs. 200 per month (USD 3 per month) was benchmarked to that charged by providers of private tutoring in the vicinity.

Third, we find that treatment effects do not vary significantly by baseline test scores, gender, or household socioeconomic status. Thus, consistent with the promise of customized technology-led instruction, the Mindspark intervention was equally effective at improving test scores for *all* students. However, while the absolute impact of Mindspark was similar at all parts of the initial test score distribution, the *relative* impact was much greater for weaker students because the 'business as usual' rate of progress in the control group was close to zero for students in the lower third of the initial test score distribution.

Fourth, using detailed electronic records of every question presented to students in the treatment group by the Mindspark program, we document that: (a) there is considerable variation in students' initial preparation for grade-appropriate work, with students enrolled in the same grade typically spanning *five to six grade levels* in their readiness; and (b) the software targets instruction very precisely to each student's learning level, and updates this targeting in response to changes in student learning. Thus, the ability of Mindspark to handle the heterogeneity in student preparedness spanning several grades appears to be an important (though not exclusive) mechanism of impact.

Fifth, Mindspark was highly cost effective. The test score value-added in the treatment group (even based on ITT estimates) was over 100% greater than that in the control group, and was achieved at a lower expenditure per student than incurred in the public schooling system. Our results are particularly striking when considered in terms of productivity per unit of time. For instance, Muralidharan (2012) finds that providing individual-level performance bonuses to teachers in India led to test score gains of 0.54σ and 0.35σ in math and language for students exposed to the program for five years. This is one of the largest effect sizes seen to date in an experimental study on education in developing countries. Yet, we estimate that Mindspark was able to achieve similar gains in one tenth the time (half a year).

Our first contribution is to the literature on computer-aided learning (CAL), where the evidence to date has often been characterized as mixed (Bulman and Fairlie, 2016). Nevertheless, our reading of the evidence suggests that some clear patterns are starting to emerge (see Appendix B for our synthesis). Hardware-focused interventions that provide computers at home or at school seem to have no positive impact on learning outcomes (Angrist and Lavy, 2002; Barrera-Osorio and Linden, 2009; Beuermann et al., 2015; Cristia et al., 2012; Malamud and Pop-Eleches, 2011).[4] Pedagogy-focused CAL programs that allow students to review grade-appropriate content at their own pace do better, but the gains are modest and range from 0.1σ to 0.2σ.[5] Finally, interventions that use technology to also personalize instruction seem to deliver substantial gains. For instance, Banerjee et al. (2007) find math

[4]These disappointing results are likely explained by the fact that hardware-focused interventions have done little to change instruction, and at times have crowded out student time for independent study.

[5]See, for example, Carrillo et al. (2010); Lai et al. (2015a, 2013, 2012); Linden (2008); Mo et al. (2014b); Barrow et al. (2009); Rouse and Krueger (2004). Anecdotal evidence suggests that pedagogy-focused CAL

test score gains of 0.47σ in two years from a CAL program that allowed some personalization. Our results finding large test-score gains from attending Mindspark centers, combined with question-level data showing how Mindspark addressed the considerable heterogeneity in student preparation are consistent with this idea, and suggest that personalization may be a key ingredient for achieving the full potential of technology-aided instruction.

Second, we contribute three key facts to the literature on education in developing countries. Prior work has posited with indirect evidence that a likely reason for low education productivity in these settings may be the combination of curricula designed for high-achieving students, increasing heterogeneity in student preparation resulting from rapid expansion of education access, and curricular instruction therefore being at a level and pace that may be too high for most students (Glewwe et al., 2009; Banerjee and Duflo, 2012; Pritchett and Beatty, 2015). We provide direct evidence in support of this hypothesis by documenting (a) large gaps between student preparation and grade-level standards that grow by grade, (b) considerable heterogeneity in student preparation in the same grade, and (c) a lack of progress in learning under the status quo for students with low initial learning levels.

Third, we contribute to the evidence on policy options to address the challenge of variation in student preparation. The most promising approaches to date have involved either explicit tracking of early grade classrooms (Duflo et al., 2011) or grouping primary school students by their level of preparation and teaching them basic skills (Banerjee et al., 2007, 2016). However, it is not clear if this approach can be extended to secondary grades where the content is more advanced and complex, and the heterogeneity in student preparation is much higher (exacerbated by "social promotion" policies in many countries). These conditions make the effective delivery of *any* curriculum challenging even for highly motivated and trained teachers, and our results suggest that personalized technology-aided instruction may be especially effective in such settings.

Finally, our results speak more broadly to the potential for new technologies to enable low-income countries to leapfrog constraints to development. For instance, Deaton (2013) documents that life expectancy in developing countries is much higher than historical levels in OECD countries at comparable stages of development, and suggests that these are likely due to new medical technologies (such as vaccinations and antibiotics) that are available now. Our results point to the possibility that technology-aided instruction could help developing countries achieve better learning outcomes by leapfrogging constraints associated with low levels of per-capita income including low teacher and parent human capital.[6]

interventions have typically focused on grade-appropriate content in response to schools' and teachers' preference for CAL software to map into the topics being covered in class and reinforce them.

[6]Examples of such technology-enabled leap-frogging from other sectors include the use of mobile telephones to circumvent the lack of formal banking systems (Jack and Suri, 2014), the use of electronic voting machines

The rest of this paper is organized as follows. Section 2 describes the intervention, sampling strategy, and randomization. Section 3 presents the data collected for this study. Section 4 discusses the empirical strategy and reports the results. Section 5 presents the cost-effectiveness analysis and discusses policy implications. Section 6 concludes.

2 Intervention and Study Design

2.1 The Mindspark CAL software

The Mindspark CAL software, developed by Educational Initiatives (a leading Indian education firm), is the central component of the program we study. The software is interactive and includes continuous student assessment alongside instructional games, videos, and activities from which students learn through explanations and feedback. Instruction in the program is designed to promote students' conceptual understanding of the material (instead of rote memorization). The Mindspark software reflects over a decade of product development and aims to leverage several of the posited channels by which education technology may improve pedagogy. We highlight the key design features of Mindspark here, and provide a more detailed description with examples for each of the points below in Appendix C.

First, it is based on an extensive corpus of *high-quality instructional materials*, featuring an item bank of over 45,000 test questions, iterated over several years of design and field testing. The design of the questions aims to reflects current research in effective pedagogy that is relevant to low-income settings, such as the use of same-language subtitling for teaching literacy (Kothari et al., 2002). Further, the software allows this material to be *delivered with uniform consistency* to individual students, thereby circumventing both limitations in teacher knowledge as well as heterogeneity in knowledge and teaching ability across teachers.

Second, the content presented to students is *adaptive*, with activities presented to each student being based on that student's performance. This adaptation is dynamic, occurring both at the beginning of the program based on a diagnostic assessment, and then with every subsequent activity completed. Thus, while the Mindspark content database is mapped into the curricular standards and learning objectives of the education system, an essential feature of Mindspark is that the content presented to students is *not* linked to the curriculum or textbook of the grade that the student is enrolled in. In other words, it enables dynamic "Teaching at the right level" for each individual student and can cater effectively to very wide heterogeneity in student learning levels that may be difficult for even highly trained and motivated teachers to achieve in a classroom setting.

Third, even students at approximately similar levels of understanding of a topic, may have very different specific areas of conceptual misunderstanding. Thus, the pedagogical approach

for better enfranchisement of illiterate citizens in democracies (Fujiwara, 2015) and the use of biometric authentication to circumvent literacy constraints to financial inclusion (Muralidharan et al., 2016).

needed to alleviate a student-specific conceptual "bottleneck" may be different across students. Mindspark aims to address this issue by using its large database of millions of student-question level observations to identify *patterns* of student errors and to classify the type of error and target *differentiated* instructional content accordingly (see Appendix C for examples). This attention to understanding patterns in student errors builds on an extensive literature in education that emphasizes the diagnostic value of error analysis in revealing the heterogeneous needs of individual students (see Radatz (1979) for a discussion). However, while the value of error analysis is well-known to education specialists, implementing it in practice is non-trivial and the use of technology sharply reduces the cost of doing so.[7]

Finally, the interactive user interface, combined with the individualization of material for each student, facilitates children's *continuous engagement* with the material. Mindspark makes limited use of instructional videos (where student attention may waver and cannot be monitored), choosing instead to instruct with steps that require students to constantly interact with the system. This approach promotes student attention and engagement, and also allows the system to provide feedback at the level of each intermediate step in solving a problem. Thus, Mindspark both promotes student engagement and sharply shortens the feedback loop between students attempting a problem and learning about their errors.

2.1.1 The Mindspark centers intervention

The Mindspark CAL software has been deployed in various settings: private and government schools, after-school instructional centers and individual subscription-based use at home. Here, we evaluate the supplementary instruction model, delivered in stand-alone Mindspark centers that target students from low-income households. Students sign up for the program by selecting a 90-minute slot, outside of school hours, which they are scheduled to attend six days per week. Typically, parents pay INR 200 (USD 3) per month to send their children to the program.[8]

Scheduled daily instruction in Mindspark centers is divided into 45 minutes of computer-based instruction and 45 minutes of supervised instructor-led group-based study. In the time allotted to the computer-based instruction, each child is assigned to a Mindspark-equipped computer with headphones that provides him/her with activities on math, Hindi and English. Two of the days of the week are supposed to be devoted to math activities, two days to Hindi, one day to English, and one day in which the child can choose the subject.

[7] The emphasis on error analysis reflects EI's long experience in conducting similar analyses and providing diagnostic feedback to teachers based on paper-and-pen tests (Muralidharan and Sundararaman, 2010). In other words, Mindspark is a product developed by *education specialists* using technology to improve productivity in implementing ideas that are known to improve the effectiveness of pedagogy.

[8] The intensity of the program, as well as the fee charged, was designed to be comparable to after-school private tutoring, typically conducted in groups of students, which is common in India. According to the 2012 India Human Development Survey, 43% of 11-17 year olds attended paid private tutoring outside of school.

The group-based instruction component includes all students in a given batch (typically between 12-15 students) and is supervised by a single instructor. Instructors are locally hired and are responsible for monitoring students when they are working on the CAL software, providing the group-based instruction, facilitating the daily operation of the centers, and encouraging attendance and retention of enrolled students.[9] Instruction in the group-based component consists of supervised homework support and review of core concepts of relative broad relevance for all children without individual customization. The intervention is, thus, a "blended learning" model which includes one-on-one computer-aided instruction alongside additional group academic support. In section 5, we account for the composite nature of the program in our interpretation of the program effects and cost effectiveness and argue that our results are likely to under-estimate the full potential of a "blended learning" model because the Mindspark centers did not optimize the use of the instructor for pedagogy.

2.2 Sample

The intervention was administered in three Mindspark centers in Delhi focused on serving low-income neighbourhoods. The sample for the study was recruited in September 2015 from five public middle schools close to the Mindspark centers. All five schools had grades 6-8, three of these schools had grade 9, and only two had grades 4-5. Three were girls-only schools and the other two were boys-only secondary schools. Therefore, our study sample has a larger share of girls in grades 6-8.

In each school, with authorization from the school principals, staff from EI and from J-PAL South Asia visited classrooms from grades 4-9 to introduce students to the Mindspark centers intervention and the study and to invite them and their parents to a scheduled demonstration at the nearby Mindspark center. Students were provided flyers to retain this information and to communicate with their parents.

At the demonstration sessions, students and their parents were introduced to the Mindspark intervention by staff from EI and basic background information was collected. Parents were told that, if their child wanted to participate in the study, he/she would need to complete a baseline assessment and that about half of the students would be chosen by lottery to receive a scholarship which would waive the usual tuition fees of INR 200 per month until February 2016 (i.e. for nearly half of the school year). Students who were not chosen by lottery were told that would be provided free access to the centers after February 2016, if they participated in an endline assessment in February 2016. However, lottery losers were not allowed to access

[9]These instructors are recruited based on two main criteria: (a) their potential to interact with children; and (b) their performance on a very basic test of math and language. However, they are not required to have completed a minimum level of education at the secondary or college level. They receive an initial training, regular refresher courses, and have access to a library of guiding documents and videos. They are paid much lower salaries than civil-service public-school teachers.

the program during the study period. These two design features helped to reduce attrition, and increase statistical power respectively.

Of the 766 students who attended the demonstration sessions, 619 completed the baseline tests and all sections of the survey, and were included in the study. About 97.5% of the study participants were enrolled in grades 6-9.[10] To assess the representativeness of our self-selected study sample (and the external validity of our results), we plot the test score distribution of study participants, and that of the full population of students in the same schools using administrative data on final exam scores in the preceding school year (2014-15) (Figure A.1). While study participants have moderately higher pre-program test scores than their peers—indicating modest positive selection on prior achievement—there is substantial common support in the range of achievement across participants and non-participants suggesting that our results are likely to extend to other students in this setting.

2.3 Randomization

The 619 participants were individually randomized into treatment and control with 305 students in the control group and 314 in the treatment group. Randomization was stratified by center-batch preferences.[11] The treatment and control groups do not differ significantly on any observable dimension at baseline (Table 1, Panel A). Of the 314 students offered a scholarship for the Mindspark program, over 80% attended the program for at least 7 days. There was, however, considerable variation in the number of days attended across students which we discuss when we present ITT and IV estimates of the main program effects in Section 4.

Of the 619 students who participated in the baseline test, 533 also attended the endline test (270 control students and 263 treatment students), yielding a mean follow-up rate of about 86%. The attrition rate was 16% in the treatment group, and 12% in the control group, but the difference was not significant at the 5% level. We also find no significant difference in the student characteristics (age, gender, SES, or baseline test scores) of those who attend both the baseline and endline test, and comprise our main study sample (Table 1, Panel B).

[10] 589 students were enrolled in grades 6-9, 15 were enrolled in grades 4-5 and, for 15 students, the enrolled grade was not reported. Our focus on Grades 6-9 reflects our funding from the JPAL Post Primary Education Initiative, which prioritized studying interventions to improve post-primary education (after fifth grade).

[11] Students were asked to provide their preferred slots for attending Mindspark centers given school timings and other responsibilities. Since demand for some slots is expectedly higher than others, we generated the highest feasible slot for each student with an aim to ensure that as many students were allocated to their first or second preference slots as possible. Randomization was then carried out within center-by-batch strata.

3 Data

3.1 Student achievement

The primary outcome of interest for this study is student test scores. Test scores were measured using paper-and-pen tests in math and Hindi prior to the randomization (September 2015, baseline) and near the end of the school year (February 2016, endline).[12] Tests were administered centrally in Mindspark centers at a common time for treatment and control students with monitoring by J-PAL staff to ensure the integrity of the assessments.

The tests were designed independently by the research team and intended to capture a wide range of student achievement. Assessment questions ranged in difficulty from "very easy" questions designed to capture primary school level competences much below grade-level to "grade-appropriate" competences such as found in international assessments.

Test questions were taken from independent assessments previously administered by high-quality research projects in India and internationally. Separate test booklets were developed for different grade levels, and across baseline and endline tests, but with substantial overlap in test items which allows for the generation of comparable test scores. Test scores were generated using Item Response Theory models to place all students on a common scale across the different grades and across baseline and endline assessments. The common scale over time allows us to characterize the *absolute* test score gains made by the control group between the two rounds of testing. Details of the test design and scoring are provided in Appendix D. The assessments performed well in capturing a wide range of achievement with very few students subject to ceiling or floor effects.

3.2 Mindspark CAL system data

The Mindspark CAL system collects detailed logs of all interactions that each student has with the software platform. This includes, for example, daily attendance, the estimated grade level of student achievement as determined by the Mindspark system, the record of each question that was presented to the child and whether he/she answered correctly, as well as other details such as time taken to answer or enter keystrokes to measure engagement with content. These data are available (only) for the treatment group for the duration of the intervention. We use these data in three ways: to describe the distribution of learning gaps relative to curricular standards in each grade at baseline; to demonstrate the personalization of instruction at the core of the Mindspark system; and to characterize the evolution of student achievement in the treatment group over the period of the treatment.

[12] It was important to test students in a pen-and-paper format, rather than computerized testing, to avoid conflating true test score gains with greater familiarization with computer technology in the treatment group.

3.3 School records

At the school level, we collected administrative records on academic test scores of all study students and their peers in the classroom as well as details of student attendance. This was collected for both the 2014-15 school year (in order to understand pre-existing differences between the study population and the non-study students in these schools) and the 2015-16 school year (to evaluate whether the treatment affected test scores on school exams).

3.4 Student data

At the time of the baseline assessment, students answered a self-administered written student survey which collected basic details about their socio-economic status, household characteristics and academic support including their attendance of private tutoring. A shorter survey of time-varying characteristics was also administered at endline. Additionally, we also phoned parents of the study participants to collect information on private tuition attendance in July 2016 based on retrospective recall for the last academic year and their opinion of the Mindspark program.

4 Results

4.1 Business-as-usual academic progress

Our first set of results use the rich data on student achievement to present a more granular characterization of the variation in learning levels within a typical classroom than has been possible so far. Specifically, the data from the Mindspark CAL system provides an assessment of the actual grade-level of each student's learning level regardless of grade enrolled in. These data are only available for the students in the treatment group. However, the information we present below comes from the initial diagnostic test, which was done immediately after randomization, and does not reflect any instruction provided by the program. The main results are presented in Figure 1, which shows the full joint distribution of the grades students were enrolled in and their assessed learning level at the start of treatment.

We highlight three main patterns. First, most children are already much below grade level competence at the very beginning of post-primary education. In grade 6, the average student is about 2.5 grades behind in math and about half a grade behind in Hindi.[13] Second, although

[13]These figures are likely to underestimate the gap relative to curricular standards at the end of primary school (grade 5), because weaker students are less likely to have made the transition to post-primary school enrolment and will not be in our sample. Also, the allocation of test questions to grade levels is more robust in math than language (where competencies are less well-delineated across grades). Thus, although most patterns across grades are similar in the two subjects, the computer system's assessment on grade-level competence of children is likely to be more reliable for math. Baseline test scores on our independent tests in both subjects are consistently higher for students assessed by the CAL program as being at a higher grade level of achievement, which helps to validate the grade-level bench-marking by the CAL program. Further details of the diagnostic test and bench-marking by the software are presented in Appendix C.

average student achievement is higher in later grades, indicating some learning over time, the slope of achievement gains (measured by the line of best fit) is flatter than the line of equality between curricular standards and actual achievement levels. This suggests that average student academic achievement is progressing at a lower rate than envisaged by the curriculum — by grade 9, students are (on average) nearly 4.5 grades behind in math and 2.5 grades behind in Hindi. Third, the figure presents a stark illustration of the very wide dispersion in achievement among students enrolled in the same grade: students in our sample span 5-6 grade levels in each grade. Characterizing this variation is essential for designing effective instructional strategies, and the CAL data allow us to do this much more clearly than the literature to date.

Our second descriptive result uses the IRT-linked student-level panel data from our independent tests to investigate heterogeneity in the 'business as usual' academic progress in the control group. Specifically, we present the value-added of test scores for the bottom, middle, and top third of the within-grade achievement distributions in our sample in both math and Hindi. We estimate the following regression (without a constant term):

$$Y_{is2} = \gamma . Y_{is1} + \alpha_j . \mathbf{terc}_{js1} + \epsilon_{i2} \tag{1}$$

where Y_{ist} is student i's test score on our independent assessment in subject s at period t, $\mathbf{terc}_{js1}$ is a vector of indicator variables for the within-grade terciles (j) of baseline achievement in the given subject s, and ϵ is the error term. [14]

Coefficients from the vector α_j may be interpreted as the *absolute* value-added in each tercile j. Such an interpretation is made feasible by the comparable measurement of test scores across baseline and endline tests on a *common scale*, which is done here by linking test scores using IRT. These coefficients are presented in Figure 2. Students at different parts of the distribution make very different progress — initially better-achieving students also have higher value-added over the period between baseline and endline. Strikingly, we cannot reject the null of *no increase* in test scores for the bottom-third in both subjects and the coefficients in both math and Hindi in this group are close to zero in absolute magnitude.

Figures 1 and 2 highlight that a vast majority of students are left behind the curriculum, and that this gap grows by grade. Further, lower-performing students appear to make no academic progress at all. To the best of our knowledge, we are the first to present direct evidence on these facts about education in developing countries. However, they are consistent with evidence on the relatively slow progress in growth of achievement in repeated cross-sections in India (see e.g. Pritchett (2013)), and with the patterns of results observed in experimental evaluations of

[14] Test scores are normalized to have a mean of zero and a standard deviation of one on the baseline test. We use robust Huber-White standard errors throughout the paper.

education interventions in developing countries in the past decade (Glewwe et al., 2009; Duflo et al., 2011; Banerjee and Duflo, 2012). Thus, although our sample is not representative and we present these results mainly to set the context for the intervention, the patterns presented here are likely to be similar to those in other developing country settings.

4.2 Main program effects

4.2.1 Intent-to-treat estimates

Figure 3 presents mean test scores in the baseline and endline assessments in math and Hindi for lottery-winners and losers. While test scores improve over time for both groups, endline test scores are significantly and substantially higher for the treatment group. We estimate intent-to-treat (ITT) treatment effects of winning the lottery using:

$$Y_{is2} = \alpha + \gamma . Y_{is1} + \beta . Treatment_i + \phi_i + \epsilon_{it} \tag{2}$$

where Y_{ist} is student i's test score in subject s at period t; $Treatment$ is an indicator variable for being a lottery-winner; ϕ_i are stratum fixed effects to reflect the randomization design; and ϵ_{it} is the error term.

We find that students who won the lottery to attend Mindspark centers scored 0.37σ higher in math and 0.23σ higher in Hindi compared to lottery losers after just 4.5 months of winning the lottery (Table 2: Cols. 1-2). In Cols. 3 and 4, we omit strata fixed effects from the regression, noting that the constant term α in this case provides an estimate of the value-added in the control group over the course of the treatment. Again, note that this interpretation is possible because the baseline and endline tests are linked to a common metric using Item Response Theory; this would not be possible if scores were normalized within grade/period as is common practice.[15] Expressing the value-added in the treatment group ($\alpha + \beta_1$) as a proportion of the control group VA (α), these results indicate that lottery-winners made twice the progress in math, and 2.5 times the progress in Hindi, compared to lottery-losers over the study period.

4.2.2 IV estimates of dose-response relationship

The ITT results in Table 2 are estimated with an average attendance of about 58% among lottery-winners (with a maximum possible attendance of 86 days).[16] These are thus likely to be an underestimate of the program effects under full compliance.

[15] Our treatment effects, however, are of very similar magnitudes (0.36σ in math and 0.21σ in Hindi) when scores are normalized using a within-grade normalization instead (Table A.5). Note also that α here is the mean of the coefficients on within-grade terciles (α_j) presented in Figure 2.

[16] About 13% of the lottery-winners attended the program for one day or less over the period of the program. The mean attendance among the rest is about 57 days, i.e 66% of the total working days for the centers over this period. The maximum attendance recorded is 84 days (97.7%). We correlated subsequent attendance in the treatment group to various baseline characteristics, the results of which are presented in Table A.1. Students from poorer backgrounds, and students with higher baseline achievement in Hindi, appear to have greater attendance but the implied magnitudes of these correlations are small. The full distribution of attendance among lottery-winners is presented in Figure A.2.

We estimate the dose-response relationship between days of Mindspark center attendance and value-added using the following regression:

$$Y_{is2} = \alpha + \gamma . Y_{is1} + \mu_1 . Attendance_i + \eta_{it} \tag{3}$$

where Y_{ist} is defined as previously, *Attendance* is the number of days a student was reported to have logged in to the Mindspark system (which is zero for all lottery-losers) and η is a stochastic error term.

We first estimate this using OLS (Table 3: Cols. 1-2). Then, since attendance may be endogenous to expected gains from the program, we instrument days attended by the random offer of a scholarship (Table 3: Cols. 3-4). Both OLS and IV estimates show a strong and significant relationship between the number of days attended and value-added over the study period in both subjects.[17]

The IV estimate above identifies the average causal response of the treatment which "captures a weighted average of causal responses to a unit change in treatment (in this case, an extra day of attending a Mindspark center), for those whose treatment status is affected by the instrument" (Angrist and Imbens, 1995). This is identified non-parametrically under minimal assumptions of IV validity (justified here by the randomized assignment of the voucher) and monotonicity. However, using this IV estimate to predict the effect of varying the number of days attended requires further assumptions about (a) the functional form of the relationship between days attended and the treatment effect and (b) the nature of heterogeneity in treatment effects across individuals.

We explore the functional form of the relationship between attendance and learning gains for treatment group students graphically in Figure 4. Value-added increases monotonically with attendance in both subjects. The relationship seems entirely linear in math and, even in Hindi, although there are some signs of diminishing returns, we cannot reject a linear dose-response relationship (see Table A.3). Our results also indicate that variation in attendance is able to account for the full extent of the ITT treatment effects; specifically, the constant term in the OLS value-added regressions in Table 3, corresponding to zero attendance in the program, is near-identical to our estimates of value-added in the control group in Table 2, both when using the full sample and when using only data on the treatment group (Cols. 5-6).

[17] We cannot formally reject the equivalence of OLS and IV estimates; for both math and Hindi, the p-value from the difference-in-Sargan test (similar to a Hausman test, but allowing for heteroskedasticity) is substantially greater than 0.1 (Cols. 3-4). The OLS model above is a lagged value-added (VA) model, which may be thought of as a dynamic treatment effects estimator relying on ignorability. The close correspondence here between the VA and IV results adds to much recent evidence that VA models typically agree closely with experimental and quasi-experimental estimates (see, for example, Chetty et al., 2014; Deming, 2014; Deming et al., 2014; Kane et al., 2013; Angrist et al., 2015; Andrabi et al., 2011; Singh, 2015, 2016).

Taken together, these patterns suggest that a linear dose-response function is a reasonable approximation in our setting . We additionally assume constant treatment effects to provide suggestive magnitudes of the treatment effect under alternative intensities of treatment.[18] Under these assumptions, which appear reasonable in this context, our results suggest that 90 days' attendance, which roughly corresponds to half a school year with 80% attendance, would lead to gains of 0.59σ in math and 0.37σ in Hindi. We extrapolate results to 90 days, rather than a full school year, to not extend predictions far outside the range of our data.

These estimates are likely to be lower-bound estimates of the productivity of Mindspark instructional time in the particular subjects since attendance here does not account for the time spent in the Mindspark centres on instruction other than Math and Hindi (in particular, instruction in English, staff trainings, parent-teacher meetings and educational excursions). In Table A.4, we present analogous IV and value-added specifications which only account for the time spent by students on either computer-aided or small-group instruction in the particular subject; using these estimates, 90 days of instructional time, split equally between the two subjects, would lead to treatment gains of 0.76 SD in math and 0.5 SD in Hindi.

4.2.3 Program impacts by learning domain

In addition to presenting impacts on a normalized summary statistic of student learning, we also present impacts of the program on specific domains of subject-level competencies. The intervention led to significant increases across all domains of test questions (Table 4). The magnitude of these effects is substantive: expressed as a proportion of the correct responses in the control group, these ITT effects range from a 12% increase on the "easiest" type of questions (arithmetic computation) to up to 38% increase on harder competences such as geometry and measurement. Similarly, in Hindi, these effects represent increases from about 7% on the easiest items (sentence completion) to up to 19% on the hardest competence (to answer questions based on interpreting and integrating ideas and information from a passage).

4.3 Heterogeneity

We investigate whether treatment effects vary by gender, socio-economic status, or initial test scores, and find no evidence of heterogeneity on these dimensions (Table 5). We further present a non-parametric representation of the ITT effect plotting kernel-weighted local mean (Nadaraya-Watson) smoothed lines, which relate absolute endline test scores to percentiles in the baseline achievement distribution, separately for the treatment and control groups (Figure 5). In both math and Hindi, the trajectory of achievement is shifted upwards for the

[18]This assumption is required for extrapolation because the average causal response is only identified over a subset of compliers, and not the full sample. In the following section, we will document that we find no significant evidence of heterogeneity in treatment effects along multiple observable dimensions. A constant treatment effects assumption also underlies our value-added specifications which, as seen in Table 3, agree closely with the IV specifications.

treatment group and significantly different from the control group trajectory. This indicates that the treatment benefited students at all parts of the achievement distribution and relatively equally.

We focus next on heterogeneity by relative position in the within-grade achievement distribution. We regress endline achievement on the baseline test score, indicator variables for treatment, for within-grade terciles at baseline, and interaction terms between the treatment variable and two terciles; the regression is estimated without a constant.We see no significant evidence of heterogeneity (see Table 6) - although the coefficient on the treatment dummy itself is strongly statistically significant, the interaction terms of treatment with the tercile at baseline are in all cases statistically indistinguishable from zero.

These results indicate that the Mindspark intervention could teach all students equally well, including those in the lowest terciles who were not making any academic progress under business-as-usual. Moreover, expressing gains from the treatment as a multiple of what students would have learnt in the absence of treatment, it is evident that the treatment effect is a larger relative effect for weaker-performing students.

4.4 Personalization

The computer-based instruction in Mindspark combines effects from multiple channels: uniformly high quality content, personalization, shorter feedback loops, and possibly a more engaging format. We cannot separately identify the effects of these channels individually. However, the detailed question-level data collected in the Mindspark system for individual students in the treatment group does allow us to examine more closely a key component of the intervention's posited theory-of-change — the delivery of personalized instruction which is able to target student preparedness precisely and update instruction appropriately.

We first present direct evidence that the Mindspark system precisely targets instructional material at an individual student's level (Figure 6). We show, separately by each grade of school enrolment, the actual grade level of a student's academic preparedness as estimated by Mindspark CAL system and the grade-level difficulty of the questions that he/she was presented in a single day.[19] Across the horizontal axis on each subgraph, we see the wide dispersion in academic preparedness within each grade, reiterating our interpretation of Figure 1. On the vertical axis, however, we see that the Mindspark system is able to precisely target instruction to preparedness and that the typical child is presented items either at their grade level or adjacent. This degree of individualization is considerably more precise than would be feasible for a single teacher to deliver to all students in a standard classroom setting.

[19] In both math and Hindi, we use data from a single day which is near the beginning of the intervention,after all students would have completed their initial assessment, and when Mindspark computer-aided instruction in the relevant subject was scheduled in all three centers.

Second, we show that the Mindspark CAL system accommodates variation, not just in the incoming levels of student preparedness, but also in their pace of learning and individual trends in student achievement. Figure 7 presents non-parametric plots of the difficulty level of the math items presented to students over the course of the intervention, documenting that the software updates its estimate of student achievement levels in real time and modifies instruction accordingly.[20] In the first figure, separate lines are plotted by the grade children are enrolled in and, in the second figure, by their initial level of achievement. As can be seen, this estimated level of difficulty increases for all groups of students continuously indicating that students were making progress regularly over time during the study period and that the Mindspark software was able to customize instruction to their increasing achievement. The individualization of the dynamic updating of content is highlighted further in Figure A.3 where we use the rich individual data on student achievement to plot similar trajectories for each child in the treatment group separately.

In summary, the Mindspark system does seem to fulfil the promise of granular customization at a level that may only be possible with either individual tutoring or perfect academic tracking but is not feasible under most current models of classroom instruction. Insights from previous work suggest that this is likely to be a key channel of impact. Together with the uniform delivery of high quality content, this highlights the potential for education technology to significantly change the delivery of instruction and improve educational productivity.

4.5 Effect on school tests

Given substantial deficits in student preparation (Figure 1), even large absolute increases in skills may not be sufficient for raising grade-level achievement.[21] This is made more likely with precise personalization of content to student levels, as students are possibly faced with little 'grade-appropriate' instruction. We investigate this directly with the CAL software data and examine the grade level of content presented by Mindspark to students in the treatment group (see Figure 8). The figure confirms our intuition: in math, very few items were administered at the level of the grade the child is enrolled in or the grade immediately below; in contrast, a substantial portion of the Hindi instruction in each grade was at grade level.

[20] We study dynamic updating only in math because the algorithms for updating are more finely developed, and differ importantly from, Hindi. In math, the software moves students to a harder question in the same competence after answering initial question(s) correctly but in Hindi the software ensures that a student has mastered all competences at a grade level before being presented questions at the next grade level.

[21] This is important to investigate separately because school tests may be high-stakes and because, in a context where many students are first-generation learners, school tests may be the most salient metric for parents to judge the academic performance of their children and decide on levels of investment into education (see e.g. Dizon-Ross (2014)).

Next, we explore the treatment effects expressed at the proportion of test questions answered correctly at grade level and at below-grade level.[22] This is presented in Table 7 for both math and Hindi. In math, consistent with little teaching at grade-level competence, we find no evidence of a treatment effect on grade level questions – the estimated coefficient on the treatment dummy variable is statistically insignificant and very close to zero in magnitude – although we find evidence of a significant and substantial treatment effect on items below grade level. In Hindi, on the other hand, where much of teh material presented by Mindspark *was* at grade level, we find that the treatment effect is significant in all regressions.

Finally, Table 8 presents the treatment effect of being offered a voucher on scores on the school exams held in March 2016.[23] Mirroring the results on grade-level items on our own tests, we find a significant increase in test scores of about 0.19σ in Hindi but no significant effect on math. We also do not find any significant effect on the other subjects (Science, Social Science or English), although coefficients are invariably positive.

4.6 Other private tutoring

We collected data from parents of the study children, using phone surveys, on whether the student attended paid extra tutoring (other than Mindspark) in any subject separately for each month from July 2015 to March 2016. Dividing this period into "pre-intervention" (July to September 2015) and "post-intervention" (October 2015 to March 2016), we test whether winning a Mindspark-voucher affected the incidence of private tutoring by estimating:

$$T_{ism} = \alpha + \phi_1.post + \phi_2.post * Treatment_i + \lambda_{is} + \epsilon_{it} \tag{4}$$

where T_{ism} is an indicator variable for whether child i attended private tutoring in subject s in month m, $Treatment$ is a binary indicator for lottery-winners and post is a binary indicator for a time period after September 2016. λ_{is} is a set of individual fixed effects. We present these results in Table A.6 and find no evidence of any differential use of private tutoring among lottery winners.

4.7 Robustness

We test the robustness of our results to attrition by modeling selection into the endline based on observed characteristics, and present inverse probability weighted treatment effects: the estimated ITT effects are substantively unchanged (Table A.7). WE also compute Lee (2002) bounds for the treatment effect: although bounds are wide, the treatment effects are always positive and significant (Table A.8).

[22]Items on our tests, which were designed to capture a wide range of achievement, were classified ex-post as belonging uniquely to a particular grade-level with the help of a curriculum expert.

[23]March is the end of the academic year in India, when students sit end-of-year exams. In Delhi, test papers are common across schools for each subject in each grade. In the regressions above, scores are standardized to have a mean of zero and a standard deviation of 1 in each grade/subject in the control group.

A further potential concern relates to teaching to the test: our independent assessments used test items from several external assessments, some of which (in the Indian setting) were designed by EI; this raises the possibility that results on our assessments are overstated due to duplication of items between our tests and the Mindspark item bank.[24] We test this by computing the treatment effect expressed as the proportion correct on items from EI assessments and items from other assessments. The ITT effects are positive, statistically significant and of similar magnitude for both sets of items in math and Hindi (Table A.9).

5 Discussion

5.1 Cost-effectiveness

We now compare the efficacy and cost-effectiveness of the Mindspark centres to other alternatives for education services available in this context.

5.1.1 Comparison with after-school private tutoring

Since the Mindspark centers program was offered after school, a natural comparison is with after-school private tutoring, which is commonplace in India and in many other developing countries. Berry and Mukherji (2016) conduct an experimental evaluation, contemporaneous to our study, of paid private tutoring with a sample of students in grades 6-8 in Delhi. The program also provided six days of instruction per week, charged INR 200 per month (which was the subsidized fee charged by Mindspark centers), and students were taught for two hours per day (25% more scheduled instruction time than the Mindspark intervention). The intervention was run by a well-respected and motivated non-governmental organization, Pratham, which has previously shown positive effects of other interventions (see, for example, Banerjee et al., 2016, 2007).

Despite similarities, this intervention differs from ours in two significant respects - instruction is delivered at grade-level curriculum and not customized to the academic preparation of the child and, secondly, the instruction is delivered in person by a tutor in groups of up to 20 students (similar to the group-based instruction in Mindspark centers). Both these features are typical of existing private tutoring market in India. At the end of a year of instruction, Berry and Mukherji (2016) find no evidence of a significant treatment effect in either math or English, the two subjects in which they, like us, administer independent assessments. Thus the best evidence available so far suggests that teacher-led group-based tutoring, delivered in similar samples and with a treatment dosage more intensive than Mindspark, was unable to deliver positive treatment effects when instruction was tied to the grade level of the child.

[24] Note, however, that the Mindspark Item Bank contains over 45,000 items and so mere duplication in the database does not imply that a student would have been presented the same item during the intervention.

5.1.2 Comparison with business-as-usual public schooling

A second possible comparison is with the productivity of government-run schools (from where the study subjects were recruited). Per-pupil monthly spending in these schools in Delhi was around INR 1500 (USD 22) in 2014-15; students spend 240 minutes per week on math and Hindi; and we estimate that the upper-bound of the value-added in these schools was 0.36σ in math and 0.15σ in Hindi over the 4.5 month study period. Specifically, this was the *total* value-added in the control group in Table 2, which also includes the effects of home inputs and private extra tutoring, and therefore over-estimates learning gains in public schools.

Using our ITT estimates, we see that Mindspark added 0.37σ in math and 0.23σ in Hindi over the same period in around 180 minutes per week on each subject. The Mindspark program, as delivered, cost about INR 1000 per student (˜USD 15) per month. This includes the costs of infrastructure, hardware and staffing as well as costs for software development. Thus, even when implemented with high fixed costs and without economies of scale, and based on 58% attendance, the Mindspark intervention delivered greater learning at lower financial and time cost than default public spending.

Steady-state costs of Mindspark at policy relevant scales are likely to be much lower since (high) fixed costs of product development have been incurred already. If implemented in government schools, at even a modest scale of 100 schools, per-pupil costs reduce to about USD 2 per month (including hardware costs but excluding rent and utilities); at a scale of 1000 schools, these reduce to less than USD 10 annually. At scale, the per-pupil costs of the software and technical support alone are expected to be below USD 2 annually, which is a small fraction of the USD 150 annual cost (over 10 months) during our pilot. The program thus has the potential to be very cost-effective at scale.

5.2 Interpreting a composite treatment effect

Our estimates of program impact combine the effect of group-based and computer-based instruction. Given the composite nature of the program, we cannot provide causal decompositions of the treatment effect into these two components. However, we think it likely that a substantial portion results directly from the CAL component. The precise customization in the computer-based instruction is not achievable by even motivated teachers for groups with wide variation in student achievement. Since students selected into batches based on individual convenience, groups for instructor-led teaching span many grade levels both in terms of the actual grade students are enrolled in and their academic preparation (See Appendix Figure A.4). This is particulary marked in comparison to other supplementary instruction programs delivered in these settings, for example Berry and Mukherji (2016), where within-group heterogeneity is partially controlled for by grouping students by enrolled grade level or academic preparation. Despite this, we have shown that the program, as delivered and accounting for *all* costs, was cost-effective in comparison to usual alternatives.

According to the implementers, the primary role of the instructor was to ensure adherence to the program, to encourage regular attendance by students and to focus on homework and examination preparation (which parents demand). Group instruction in the centres focused mostly on homework support or the revision of primary school level foundational skills but was not customized for individual students. Our results thus reflect the effectiveness of a "blended learning" model with a particularly well-developed CAL component but with the instructor-led component not fully optimized for effective learning. We interpret our results, therefore, as underestimates of the full potential of this class of composite interventions.

5.3 Implications for policy and research

Our results have broad relevance for several current issues in education policy in developing countries. First, policy makers in many developing countries (including India) have already demonstrated much interest in using technology in education.[25] Unfortunately, most current attention in this area is restricted to purchasing hardware with very little focus on how technology should be deployed for effective pedagogy. Given that substantial financial resources are being dedicated to the purchase of computer hardware, which evidence suggests is unlikely to improve student learning by itself, the marginal cost of deploying the Mindspark software would be particularly low and the potential for improving student achievement is high. Such a deployment would offer a natural opportunity to test impacts at a larger scale.

Second, while there is a contentious ongoing debate across countries around the potential trade-offs between academic standards and socially-equitable automatic promotion, there is much less evidence on how to teach effectively in such settings with severe learning deficits and wide within-grade variation.[26] Our results offer insight in this area. In particular, we are able to show that well-designed interventions that target the actual preparation levels of children may allow effective teaching, even with very wide within-grade dispersion in academic achievement. The promise of education technology is to enable such personalization at scale even where ability-based academic tracking within the grade may not be feasible as a means of handling heterogeneity.

Third, we also speak to the broader (mis-)orientation of the Indian education system, which is often thought to cater to the top-end of the distribution and focus far more on screening than teaching all students effectively.[27] Over-ambitious curricula, the neglect of weak performance

[25]For instance, various state governments in India have distributed free laptops to students in recent years. Many governments have also invested in the creation of computer labs in school (such as the Adarsh schools in Rajasthan). And the emphasis on technology in education has also featured in large national level policy approaches such as the Digital India initiative of the current Union government.

[26]For examples of empirical work evaluating the effects of social promotion, see for example, Jacob and Lefgren (2004) in the US, Manacorda (2012) in Uruguay and Koppensteiner (2014) in Brazil.

[27]For an indirect illustration of this phenomenon, see Das and Zajonc (2010) who document that although the top 5% of the Indian distribution perform comparably with their international peers, the rest of the distribution performs much worse. They estimate that the Indian distribution of student achievement is the

in most of the distribution, and the focus on a small well-performing minority are, plausibly, all symptoms of misaligned priorities. Our contribution is to demonstrate directly that value-added in the control group, receiving business-as-usual inputs, is substantially lower for weaker students; in the control group, we cannot reject that students in the bottom third of the distribution made no progress in both math and Hindi over the school year. While not a comprehensive solution, our results indicate that technology-aided instruction can effectively reach students left behind by the current orientation of the education system.

The orientation of the schooling system also relates directly to the question of why, if the program is so successful, is it not adopted more enthusiastically by households? There is clearly a large market for supplementary tutoring in India at all levels of education. But the Mindspark centers themselves did not generate adequate take-up without the scholarships and, indeed, the centers all closed down soon after the conclusion of our experiment in the face of low demand and under-subscription. Low take-up outside of the evaluation may potentially reflect that parents are not well-informed of the efficacy of the intervention.[28] Alternatively, it is possible parents are unwilling to pay for instruction that improves learning outcomes but may not improve, at this late stage, performance in high-stakes matriculation exams and the chance for securing coveted formal sector employment.Understanding determinants of household willingness-to-pay for educational inputs and services remains an important area for further investigation. While reallocating public expenditure away from unproductive uses (see, for example de Ree et al., 2015) is likely to be welfare-improving, substantial gains may also be realized through better allocations of household resources. [29]

Finally, our results hold out two important messages for the design of student assessments in the economics of education in developing countries. First, student assessments need to be sensitive to the level and the wide range in achievement in these settings: in particular, tests which constitute only items at grade-appropriate levels of difficulty may miss substantial

second most unequal distribution for which data is available (behind only South Africa, which has a particular history of inequality).

[28] There is some suggestive evidence that this may be the case. Students and parents did respond to our (low-intensity) recruitment drives in schools. If they turned up to the demonstration sessions, they were likely to also enrol in the study. And if they won the lottery, they were substantially likely to enrol in the intervention subsequently. In this situation, experimental evaluations increasing the information available to parents may well be worth fielding, as in other domains of inaccurate information available to parents (see, for example, Dizon-Ross, 2014; Jensen, 2010).

[29] Encouraging such reoptimization by households will, however, require a more detailed understanding of how households value the causal contributions of educational services to learning. For recent examples of such exercises in the context of private schooling, see e.g. Carneiro et al. (2016) and Bau (2016). That parents and students are willing to invest in response to changes in their perceived economic returns has been documented by several recent studies (see, for example, Munshi and Rosenzweig, 2006; Jensen, 2010, 2012). In this context, experiments which vary the price of the intervention while providing information on learning gains may isolate the willingness to pay for skill acquisition which would provide valuable insights. For examples of such experiments, see Dupas and Miguel (2016) on health and Berry and Mukherji (2016) in education.

improvements in the achievement of students who are several grade levels behind. This is evident in the contrast between our results on independently-designed tests and in the grade-appropriate school tests: while we see substantial effects in both math and Hindi in our independent tests (which were designed to capture a wide range of achievement), in the school tests we only see effects on Hindi where students were closer to curriculum-expected skill levels. More generally, it reinforces the need to examine carefully the content and appropriateness of test domains, and to disaggregate treatment effects on these, in addition to presenting standardized z-score estimates. Second, we highlight the importance of comparable measurement across grades and rounds of testing: our ability to characterize *absolute* value-added over time in the control group, including by different terciles, and to be able to compare our treatment effect estimates to "business-as-usual" productivity rely importantly on our comparable measurement and test design (which are explained in greater detail in the Appendix). This is particularly important for comparing cost-effectiveness to alternative means of providing educational services, which is essential when evaluating interventions (such as ours) where inputs are provided in addition to business-as-usual.

6 Conclusions

We have presented an experimental evaluation of a technology-led supplementary instruction program targeted at improving learning outcomes in post-primary grades. We document substantial positive effects of the program on both math and language test scores and show that the program is very cost-effective both in terms of time and money. The program is effective at teaching students at all levels of prior achievement, including students in the bottom-third of the within-grade distribution who are left behind by business-as-usual instruction. This is consistent with the promise of computer-aided instruction to be able to teach *all* students effectively. Using detailed information on the material presented to students in the treatment group, we demonstrate the program was successful at targeting instruction precisely to the students' level of achievement and in handling wide variation in the academic preparation in the same grade. In Hindi, where initial deficits from curricular standards were assessed to be less severe and the computer program presented material at curricular levels, we also document strongly significant impacts on grade-level tests administered in school.

These substantial effects reflect, in our opinion, the ability of the Mindspark program to target multiple constraints that lead to the low productivity of instructional time in Indian schools. The high quality of content, combined with effective delivery and interface, may help circumvent constraints of teacher human capital and motivation. Personalized instruction makes it possible to accommodate large deficits in initial student preparation and wide variation within a single grade. Algorithms for analyzing patterns of student errors and providing differentiated feedback and follow up content that is administered in real-time, allows for feedback that is more relevant and much more frequent. These features all reflect continuous and iterative program development over a long period of more than a decade.

These effects may plausibly be increased even further with better design. It is possible that in-school settings may have greater adherence to the program in terms of attendance. It may also be possible to improve the effectiveness of teacher-led instruction in a 'blended learning' environment by using the extensive information on student-performance to better guide teacher effort in the classroom. This "big data" on student achievement also offers much potential of its own. Foremost, it can provide much more granular insight into the process of student learning than has been possible thus far – this may be used to further optimize the delivery of instruction in the program and, plausibly, also for the delivery of classroom instruction. Finally, the detailed and continuous measures of effort input by the students can be used directly to reward students, with potentially large gains in student achievement.[30]

However, there are also several reasons to be cautious in extrapolating the success of the program more broadly. The intervention, as evaluated in this paper, was delivered at a modest scale of a few centers in Delhi and delivered with high fidelity on part of the providers. Such fidelity may not be possible when implementing at scale. Additional issues relate to the mode of delivery. We have only evaluated Mindspark in after-school centers and it is plausible that the effectiveness of the system may vary significantly based on whether it is implemented in-school or out-of-school; whether it is supplementary to current classroom instruction or substitutes away current instructional time; and whether it is delivered without supervision, under the supervision of current teachers or under the supervision by someone else (e.g. the Mindspark center staff). Identifying the most effective modes of delivery for the program is likely to be a useful avenue of future enquiry.[31]

Overall, our present study is best regarded as an efficacy trial documenting that well-designed and implemented technology-enabled learning programs can produce large gains in student test scores in a relatively short period of time. Our results suggest that the promise of technology to sharply improve productivity in the delivery of education is very real, and that there may be large returns to further research on effective ways of delivering these benefits at a larger scale. This promises to an exciting area for future research.

[30]Direct evidence that this may be possible is provided by Hirshleifer (2015) who uses data from a (different) computer-aided instruction intervention to reward student effort and documents large effects of 0.57σ. See also Behrman et al. (2015) who document that incentives to students were most effective when aligned with the incentives of teachers; technology-aided programs may make student incentives more productive by decreasing the salience of teacher incentives by providing uniformly high-quality content.

[31]A useful example of such work has been the literature that followed the documenting of the efficacy of unqualified local volunteers, who were targeting instruction to students' achievement levels, in raising achievement in primary schools in two Indian cities by Banerjee et al. (2007). Subsequent studies have looked at the effectiveness of this pedagogical approach of "Teaching at the Right Level" in summer camps, in government schools and delivered alternately by school teachers and by other volunteers (Banerjee et al., 2016). The approach is now being extended at scale in multiple state education systems.

References

Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc, "Do value-added estimates add value? Accounting for learning dynamics," *American Economic Journal: Applied Economics*, 2011, *3* (3), 29–54.

Angrist, Joshua and Victor Lavy, "New evidence on classroom computers and pupil learning," *The Economic Journal*, 2002, *112* (482), 735–765.

Angrist, Joshua D and Guido W Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 1995, *90* (430), 431–442.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters, "Leveraging lotteries for school value-added: Testing and estimation," 2015. (NBER Working Paper No. 21748). Cambridge, MA: National Bureau of Economic Research (NBER).

ASER, "Annual status of education report (rural) 2014," 2014. New Delhi, India: ASER Centre.

Banerjee, A. V., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton, "Mainstreaming an effective intervention: Evidence from randomized evaluations of Teaching at the Right Level in India," 2016. *Journal of Economic Perspectives*, forthcoming.

Banerjee, Abhijit and Esther Duflo, *Poor economics: A radical rethinking of the way to fight global poverty*, New York, NY: Public Affairs, 2012.

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

Barrera-Osorio, Felipe and Leigh L Linden, "The use and misuse of computers in education: evidence from a randomized experiment in Colombia," 2009. (World Bank Policy Research Working Paper No. 4836.) Washington, DC: The World Bank.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse, "Technology's edge: The educational benefits of computer-aided instruction," *American Economic Journal: Economic Policy*, 2009, *1* (1), 52–74.

Bau, Natalie, "School competition and product differentiation," Technical Report, University of Toronto 2016.

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin, "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools," *Journal of Political Economy*, 2015, *123* (2), 325–364.

Berry, J. and P. Mukherji, "Pricing of private education in urban India: Demand, use and impact," 2016. *Unpublished manuscript.* Ithaca, NY: Cornell University.

Beuermann, Diether W, Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo, "One Laptop per Child at home: Short-term impacts from a randomized experiment in Peru," *American Economic Journal: Applied Economics*, 2015, *7* (2), 53–80.

Borman, G. D., J. G. Benson, and L. Overman, "A randomized field trial of the Fast ForWord Language computer-based training program," *Educational Evaluation and Policy Analysis*, 2009, *31* (1), 82–106.

Bosworth, B., "The Internet and the university," 2005. In Devlin, M., Larson, R. & Meyerson, J. (eds.) *Productivity in education and the growing gap with service industries*. Cambridge, MA: Forum for the Future of Higher Education & Boulder, CO: EDUCAUSE.

Bulman, G. and R.W. Fairlie, "Technology and Education: Computers, Software and the Internet," in Eric Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Elsevier, 2016, pp. 239–280.

Campuzano, L., M. Dynarski, R. Agodini, K. Rall, and A. Pendleton, "Effectiveness of reading and mathematics software products: Findings from two student cohorts," 2009. *Unpublished manuscript*. Washington, DC: Mathematica Policy Research.

Carneiro, Pedro, Jishnu Das, and Hugo Reis, "The Value of Private Schools: Evidence from Pakistan," Technical Report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies 2016.

Carrillo, Paul E, Mercedes Onofa, and Juan Ponce, "Information technology and student achievement: Evidence from a randomized experiment in Ecuador," 2010. (IDB Working Paper No. IDB-WP-223). Washington, DC: Inter-American Development Bank.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *The American Economic Review*, 2014, *104* (9), 2593–2632.

Cohen, Jessica and Pascaline Dupas, "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment," *The Quarterly Journal of Economics*, 2010, *125* (1), 1–45.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín, "Technology and child development: Evidence from the One Laptop per Child program," 2012. (IDB Working Paper No. IDB-WP-304). Washington, DC: Inter-American Development Bank.

Das, Jishnu and Tristan Zajonc, "India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement," *Journal of Development Economics*, 2010, *92* (2), 175–187.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for nothing? Experimental evidence on the impact of an unconditional teacher salary increase on student performance in Indonesia," 2015. (NBER Working Paper No. 21806). Cambridge, MA: National Bureau of Economic Research (NBER).

Deaton, A., *The great escape: Health, wealth, and the origins of inequality*, Princeton, NJ: Princeton University Press, 2013.

Deming, David J., "Using school choice lotteries to test measures of school effectiveness," *American Economic Review*, 2014, *104* (5), 406–11.

— , **Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger**, "School choice, school quality, and postsecondary attainment," *American Economic Review*, 2014, *104* (3), 991–1013.

Dizon-Ross, Rebecca, "Parents' perceptions and children's education: Experimental evidence from Malawi," 2014. *Unpublished manuscript.* Cambridge, MA: Massachusetts Institute of Technology (MIT).

Duflo, E., P. Dupas, and M. Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 2011, *101*, 1739–1774.

Dupas, P. and E. Miguel, "Impacts and determinants of health levels in low-income countries," 2016. (NBER Working Paper No. 22235). Cambridge, MA: National Bureau of Economic Research (NBER).

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, B. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, and W. Sussex, "Effectiveness of reading and mathematics software products: Findings from the first student cohort," 2007. *Unpublished manuscript.* Washington, DC: Mathematica Policy Research.

Fairlie, R. W. and J. Robinson, "Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren," *American Economic Journal: Applied Economics*, 2013, *5* (3), 211–240.

Fujiwara, Thomas, "Voting technology, political responsiveness, and infant health: Evidence from Brazil," *Econometrica*, 2015, *83* (2), 423–464.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin, "Many children left behind? Textbooks and test scores in Kenya," *American Economic Journal: Applied Economics*, 2009, *1* (1), 112–135.

Goolsbee, Austan and Jonathan Guryan, "The impact of Internet subsidies in public schools," *The Review of Economics and Statistics*, 2006, *88* (2), 336–347.

Hirshleifer, Sarojini, "Incentives for effort or outputs? A field experiment to improve student performance," 2015. *Unpublished manuscript.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Jack, W. and T. Suri, "Risk sharing and transactions costs: Evidence from Kenya's mobile money revolution," *The American Economic Review*, 2014, *104* (1), 183–223.

Jacob, Brian A and Lars Lefgren, "Remedial education and student achievement: A regression-discontinuity analysis," *Review of economics and statistics*, 2004, *86* (1), 226–244.

Jensen, Robert, "The (perceived) returns to education and the demand for schooling," *The Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

— , "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India," *The Quarterly Journal of Economics*, 2012, *127*, 753–792.

Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have we identified effective teachers? Validating measures of effective teaching using random assignment," 2013. Seattle, WA: Bill & Melinda Gates Foundation.

Khan, Salman, *The one world schoolhouse: Education reimagined*, Twelve, 2012.

Koppensteiner, Martin Foureaux, "Automatic grade promotion and student performance: Evidence from Brazil," *Journal of Development Economics*, 2014, *107*, 277–290.

Kothari, Brij, Avinash Pandey, and Amita R Chudgar, "Reading out of the ?idiot box?: Same-language subtitling on television in India," *Information Technologies & International Development*, 2004, *2* (1), pp–23.

_ , Joe Takeda, Ashok Joshi, and Avinash Pandey, "Same language subtitling: a butterfly for literacy?," *International Journal of Lifelong Education*, 2002, *21* (1), 55–66.

Lai, Fang, Linxiu Zhang, Qinghe Qu, Xiao Hu, Yaojiang Shi, Matthew Boswell, and Scott Rozelle, "Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai, China," 2012. (REAP Working Paper No. 237). Rural Education Action Program (REAP). Stanford, CA.

_ , _ , Xiao Hu, Qinghe Qu, Yaojiang Shi, Yajie Qiao, Matthew Boswell, and Scott Rozelle, "Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi," *Journal of Development Effectiveness*, 2013, *52* (2), 208–231.

_ , Renfu Luo, Linxiu Zhang, and Scott Huang Xinzhe Rozelle, "Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing," *Economics of Education*, 2015, *47*, 34–48.

_ , _ , _ , Xinzhe Huang, and Scott Rozelle, "Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing," *Economics of Education Review*, 2015, *47*, 34–48.

Lee, D. S., "Trimming for bounds on treatment effects with missing outcomes," 2002. (NBER Working Paper No. 277). Cambridge, MA: National Bureau of Economic Research (NBER).

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek, and Dinand Webbink, "The effect of extra funding for disadvantaged pupils on achievement," *The Review of Economics and Statistics*, 2007, *89* (4), 721–736.

Linden, L. L., "Complement or substitute? The effect of technology on student achievement in India," 2008. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL). Cambridge, MA.

Machin, Stephen, Sandra McNally, and Olmo Silva, "New technology in schools: Is there a payoff?," *The Economic Journal*, 2007, *117* (522), 1145–1167.

Malamud, Ofer and C. Pop-Eleches, "Home computer use and the development of human capital," *The Quarterly Journal of Economics*, 2011, *126*, 987–1027.

Manacorda, Marco, “The cost of grade retention,” *Review of Economics and Statistics*, 2012, *94* (2), 596–606.

Mead, R., “Learn different: Silicon Valley disrupts education,” 2016. *The New Yorker*. March 8, 2016.

Mo, Di, Johan Swinnen, Linxiu Zhang, Hongmei Yi, Qinghe Qu, Matthew Boswell, and Scott Rozelle, “Can one-to-one computing narrow the digital divide and the educational gap in China? The case of Beijing migrant schools,” *World development*, 2013, *46*, 14–29.

_ **, L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell, and S. Rozelle**, “The persistence of gains in learning from computer assisted learning (CAL): Evidence from a randomized experiment in rural schools in Shaanxi province in China,” 2014. *Unpublished manuscript.* Stanford, CA: Rural Education Action Program (REAP).

_ **, Linxiu Zhang, Renfu Luo, Qinghe Qu, Weiming Huang, Jiafu Wang, Yajie Qiao, Matthew Boswell, and Scott Rozelle**, “Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi,” *Journal of Development Effectiveness*, 2014, *6*, 300–323.

_ **, Yu Bai, Matthew Boswell, and Scott Rozelle**, “Evaluating the effectiveness of computers as tutors in China,” 2016.

Morgan, P. and S. Ritter, “An experimental study of the effects of Cognitive Tutor Algebra I on student knowledge and attitude,” 2002. Pittsburg, PA: Carnegie Learning.

Munshi, Kaivan and Mark Rosenzweig, “Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy,” *The American Economic Review*, 2006, *96* (4), 1225–1252.

Muralidharan, K., “Priorities for primary education policy in India’s 12th five-year plan,” *India Policy Forum 2012-13*, 2013, *9*, 1–46.

_ **and V. Sundararaman**, “The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India,” *The Economic Journal*, 2010, *120* (F187-F203).

_ **, P. Niehaus, and S. Sukhtankar**, “Building state capacity: Evidence from biometric smartcards in India,” *American Economic Review*, 2016, *106* (10), 2895–2929.

Muralidharan, Karthik, “Long-term effects of teacher performance pay: Experimental evidence from India,” 2012. *Unpublished manuscript.* San Diego, CA: University of California, San Diego.

Murphy, R., W. Penuel, B. Means, C. Korbak, and A. Whaley, “E-DESK: A review of recent evidence on the effectiveness of discrete educational software,” 2001. *Unpublished manuscript.* Menlo Park, CA: SRI International.

Negroponte, Nicholas, Walter Bender, Antonio Battro, and David Cavallo, “One laptop per child,” in “Keynote address at National Educational Computing Conference in San Diego, CA. Retrieved April,” Vol. 5 2006, p. 2007.

Pearson, P.D., R.E. Ferdig, R.L. Blomeyer Jr., and J. Moran, "The effects of technology on reading performance in the middle-school grades: A meta-analysis with recommendations for policy," 2005. *Unpublished manuscript.* Naperville, IL: Learning Point Associates.

Pritchett, Lant, *The rebirth of education: Schooling ain't learning*, Washington, DC: Center for Global Development, 2013.

_ **and Amanda Beatty**, "Slow down, you're going too fast: Matching curricula to student skill levels," *International Journal of Educational Development*, 2015, *40*, 276–288.

Radatz, Hendrik, "Error analysis in mathematics education," *Journal for Research in mathematics Education*, 1979, pp. 163–172.

Rockoff, Jonah E, "Evaluation report on the School of One i3 expansion," 2015. *Unpublished manuscript.* New York, NY: Columbia University.

Rouse, Cecilia Elena and Alan B Krueger, "Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program," *Economics of Education Review*, 2004, *23* (4), 323–338.

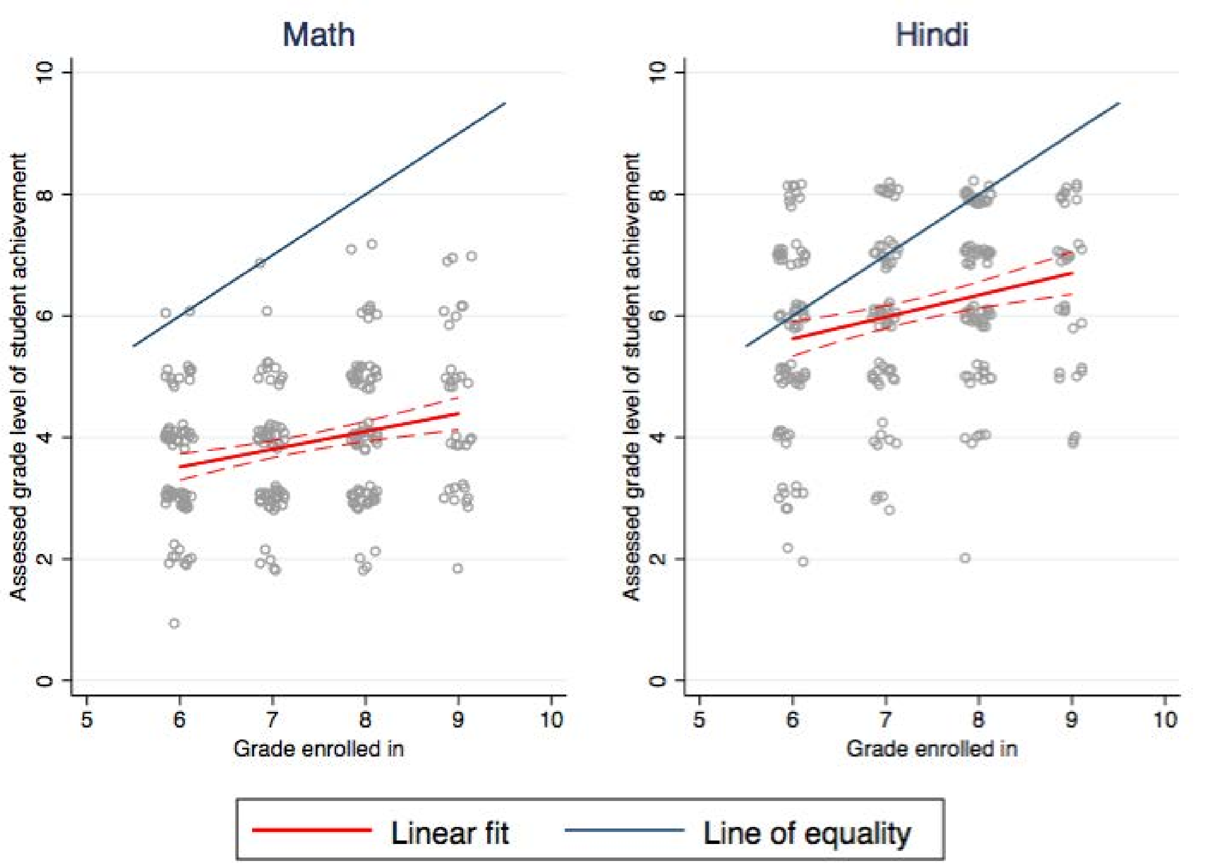
Singh, Abhijeet, "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages," *Journal of Development Economics*, 2015, *113*, 16–32.

_ , "Learning more with every year: School year productivity and international learning divergence," 2016. *Unpublished manuscript.* London, UK: University College London.

Waxman, H.C., M.-F. Lin, and G.M. Michko, "A meta-analysis of the effectiveness of teaching and learning with technology on student outcomes," 2003. *Unpublished manuscript.* CambridgeNaperville, IL: Learning Point Associates.

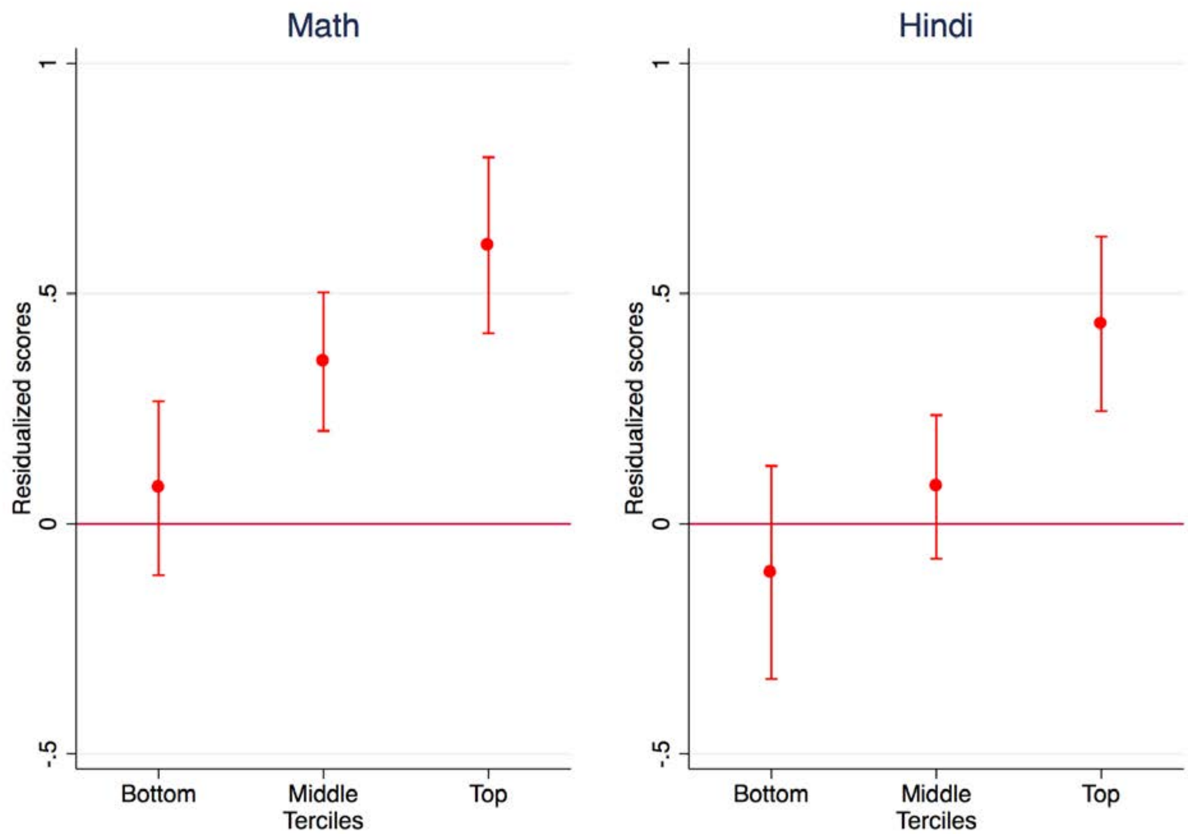
Wise, B. W. and R. K. Olson, "Computer-based phonological awareness and reading instruction," *Annals of Dyslexia*, 1995, *45*, 99–122.

Figure 1: Assessed levels of student achievement vs. current grade enrolled in school



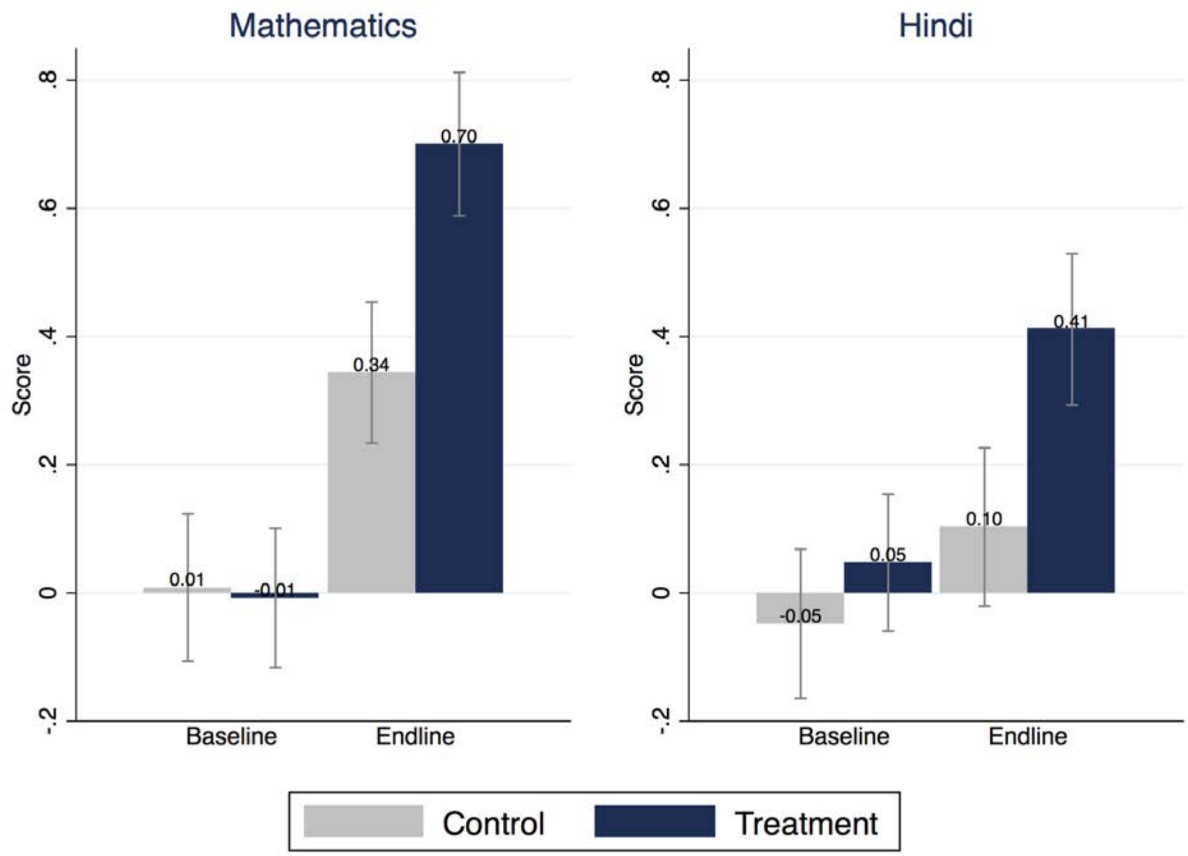
Note: This figure shows, for treatment group, the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. In both subjects, it shows three main patterns: (a) there is a general deficit between average attainment and grade-expected norms; (b) this deficit is larger in later grades and (c) within each grade, there is a wide dispersion of student achievement.

Figure 2: Business-as-usual progress in learning



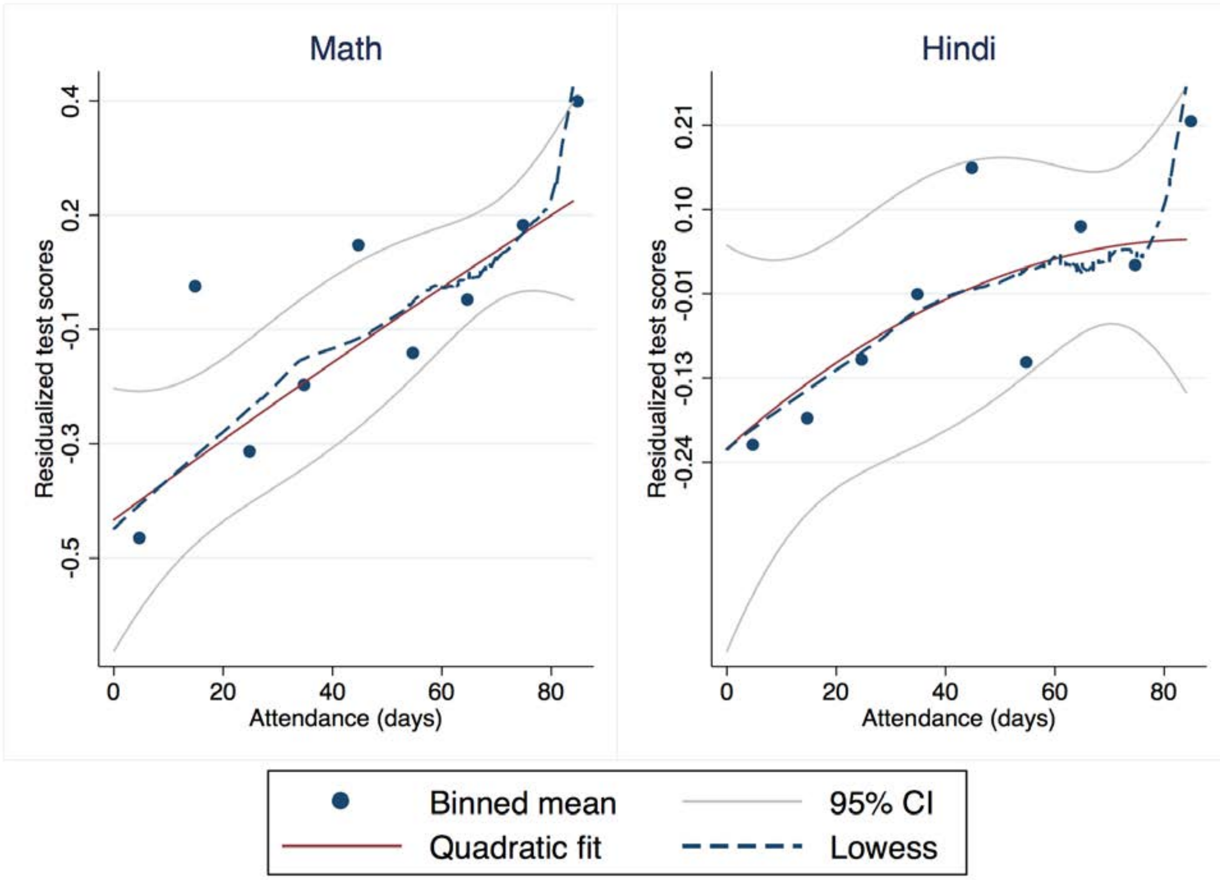
Note: This figure shows the value-added in the control group for students in different terciles of the within-grade achievement distribution. Value-added is measured on our independently-administered tests at baseline and endline tests in September 2015 and February 2016 respectively.

Figure 3: Mean difference in test scores between lottery winners and losers



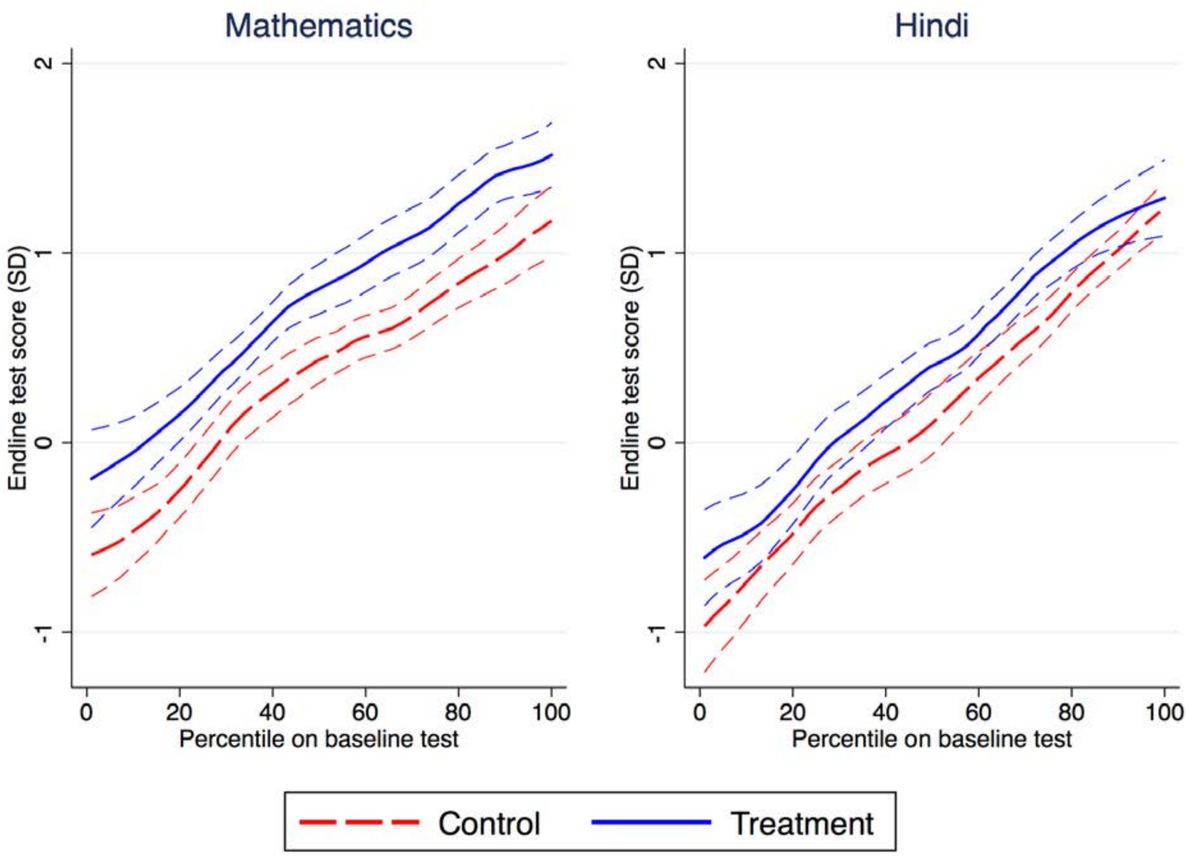
Note: This figure shows mean of test scores, normalized with reference to baseline, across treatment and control groups in the two rounds of testing with 95% confidence intervals. Test scores were linked within-subject through IRT models, pooling across grades and across baseline and endline, and are normalized to have a mean of zero and a standard deviation of one in the baseline. Whereas baseline test scores were balanced between lottery-winners and lottery-losers, endline scores are significantly higher for the treatment group.

Figure 4: Dose response relationship



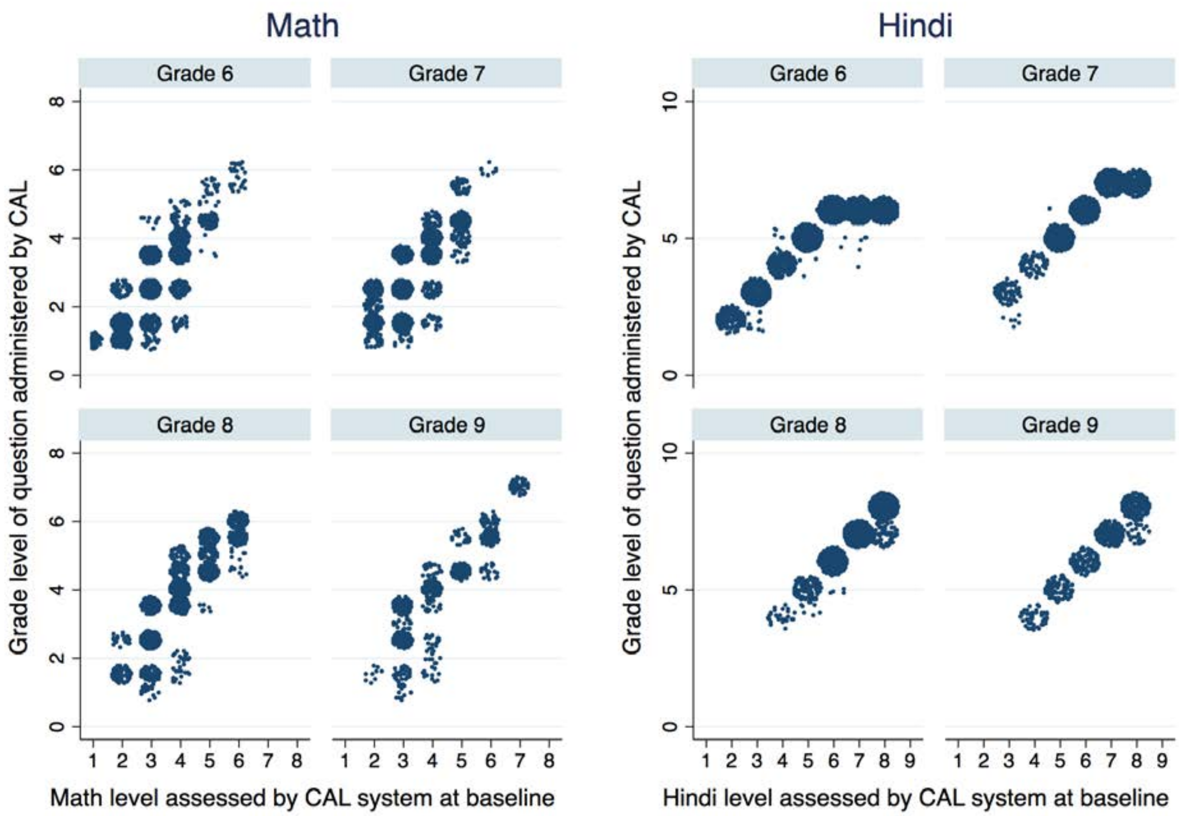
Note: This figure explores the relationship between value-added and attendance in the Mindspark program among the lottery-winners. It presents the mean value-added in bins of attendance along with a quadratic fit and a lowess smoothed non-parametric plot.

Figure 5: Non-parametric investigation of treatment effects by baseline percentiles



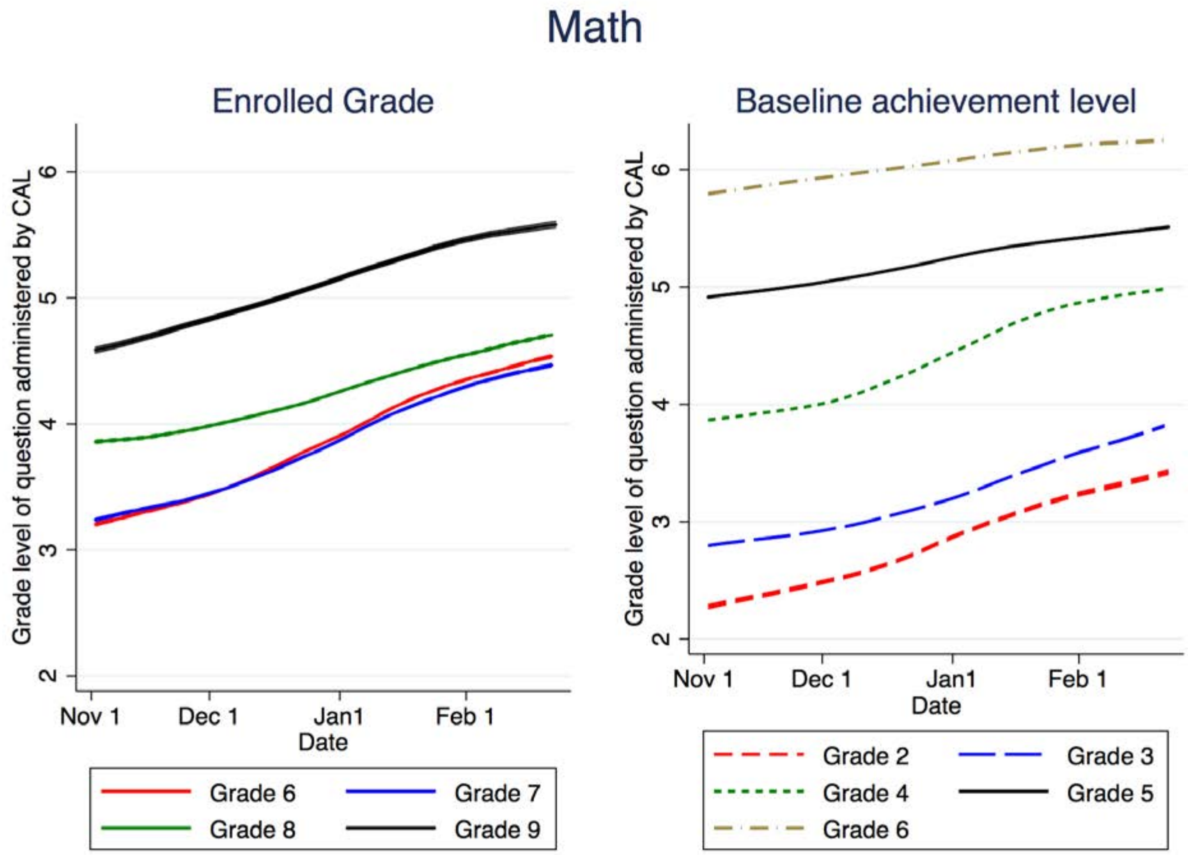
Note: The figures present kernel-weighted local mean smoothed plots which relate endline test scores to percentiles in the baseline achievement, separately for the treatment and control groups, alongside 95% confidence intervals. At all percentiles of baseline achievement, treatment group students see larger gains over the study period than the control group, with no strong evidence of differential absolute magnitudes of gains across the distribution.

Figure 6: Precise customization of instruction by the Mindspark CAL program



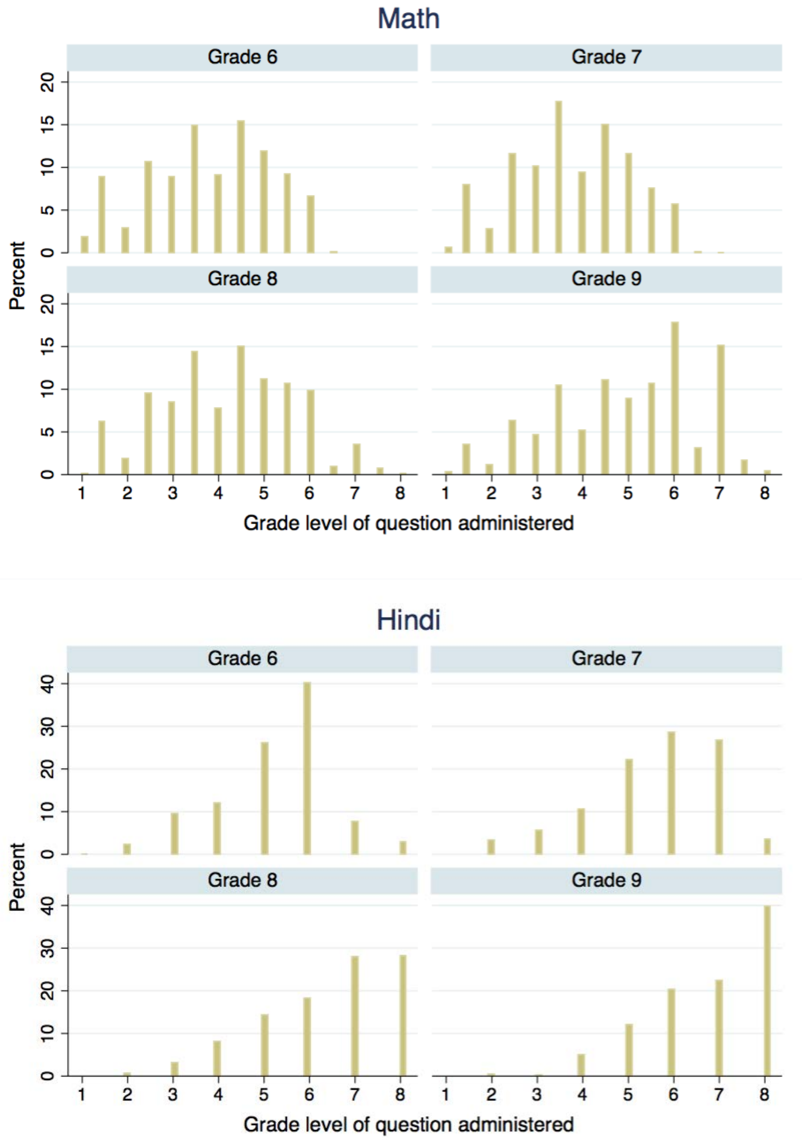
Note: This figure shows, for treatment group, the grade level of questions administered by the computer adaptive system to students on a single day near the beginning of the intervention. In each grade of enrolment, actual level of student attainment estimated by the CAL software differs widely; this wide range is covered through the customization of instructional content by the CAL software.

Figure 7: Dynamic updating and individualization of content in Mindspark



Note: This figure shows kernel-weighted local mean smoothed lines relating the level of difficulty of the math questions administered to students in the treatment group with the date of administration. The left panel presents separate lines by the actual grade of enrolment. The right panel presents separate lines by the level of achievement assessed at baseline by the CAL software. Please note 95% confidence intervals are plotted as well but, given the large data at our disposal, estimates are very precise and the confidence intervals are narrow enough to not be visually discernible.

Figure 8: Distribution of questions administered by Mindspark CAL system



Note: The two panels above show the distribution, by grade-level, of the questions that were administered by the Mindspark CAL system over the duration of treatment in both math and Hindi. Note that in math, students received very few questions at the level of the grade they are enrolled in; this reflects the system's diagnosis of their actual learning levels. In Hindi, by contrast, students received a significant portion of instruction at grade-level competence which is consistent with the initial deficits in achievement in Hindi being substantially smaller than in math (see Fig. 1).

Table 1: Sample descriptives and balance on observables

	Mean (treatment)	Mean (control)	Difference	SE	N (treatment)	N (control)
			Panel A: All students in the baseline sample			
Demographic characteristics						
Female	0.76	0.76	0.00	0.03	314	305
Age (years)	12.68	12.48	0.20	0.13	306	296
SES index	0.00	0.05	-0.05	0.14	314	305
Grade in school						
Grade 4	0.01	0.01	-0.00	0.01	305	299
Grade 5	0.01	0.02	-0.01	0.01	305	299
Grade 6	0.27	0.30	-0.04	0.04	305	299
Grade 7	0.26	0.26	0.00	0.04	305	299
Grade 8	0.30	0.28	0.02	0.04	305	299
Grade 9	0.15	0.13	0.02	0.03	305	299
Baseline test scores						
Math	-0.01	0.01	-0.02	0.08	313	304
Hindi	0.05	-0.05	0.10	0.08	312	305
Present at endline	0.838	0.885	0.048*	0.028	314	305
			Panel B: Only students present in Endline			
Demographic characteristics						
Female	0.77	0.76	0.01	0.04	263	270
Age (years)	12.60	12.46	0.13	0.14	257	263
SES index	-0.10	0.04	-0.14	0.14	263	270
Grade in school						
Grade 4	0.01	0.01	-0.00	0.01	255	266
Grade 5	0.01	0.02	-0.01	0.01	255	266
Grade 6	0.29	0.31	-0.02	0.04	255	266
Grade 7	0.25	0.25	0.00	0.04	255	266
Grade 8	0.30	0.29	0.02	0.04	255	266
Grade 9	0.14	0.12	0.02	0.03	255	266
Baseline test scores						
Math	-0.03	-0.02	-0.02	0.09	262	269
Hindi	0.06	-0.07	0.13	0.08	263	270

Note: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Treatment and control here refer to groups who were randomly assigned to receive an offer of Mindspark scholarship till March 2016. Variables used in this table are from the baseline data collection in September 2015. The data collection consisted of two parts: (a) a self-administered student survey, from which demographic characteristics, details of schooling and private tutoring are taken and (b) assessment of skills in math and Hindi, administered using pen-and-paper tests. Tests were designed to cover wide ranges of achievement and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline. The SES index refers to a wealth index generated using the first factor from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household.

Table 2: Intent-to-treat (ITT) Effects in a regression framework

	(1)	(2)	(3)	(4)
	Dep var: Standardized IRT scores (endline)			
	Math	Hindi	Math	Hindi
Treatment	0.36***	0.22***	0.36***	0.22***
	(0.063)	(0.076)	(0.062)	(0.064)
Baseline score	0.54***	0.67***	0.55***	0.69***
	(0.047)	(0.034)	(0.039)	(0.039)
Constant	0.36***	0.15***	0.36***	0.15***
	(0.031)	(0.038)	(0.043)	(0.045)
Strata fixed effects	Y	Y	N	N
Observations	529	533	529	533
R-squared	0.392	0.451	0.392	0.465

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. The SES index refers to a wealth index generated using the first factor from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household. Tests in both math and Hindi were designed to cover wide ranges of achievement and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline.

Table 3: Dose-response of Mindspark attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Standardized IRT scores (endline)					
	OLS VA (full sample)		IV models (full sample)		OLS VA (Treatment group)	
VARIABLES	Math	Hindi	Math	Hindi	Math	Hindi
Attendance (days)	0.0068***	0.0037***	0.0065***	0.0040***	0.0075***	0.0033*
	(0.00087)	(0.00090)	(0.0011)	(0.0011)	(0.0018)	(0.0020)
Baseline score	0.54***	0.69***	0.53***	0.67***	0.57***	0.68***
	(0.039)	(0.039)	(0.036)	(0.037)	(0.062)	(0.056)
Constant	0.35***	0.16***			0.31***	0.18
	(0.040)	(0.042)			(0.12)	(0.13)
Observations	529	533	529	533	261	263
R-squared	0.413	0.468	0.422	0.460	0.413	0.429
Angrist-Pischke F-statistic for weak instrument			1238	1256		
Diff-in-Sargan statistic for exogeneity (p-value)			0.26	0.65		
Extrapolated estimates of 90 days' treatment (SD)	0.612	0.333	0.585	0.36	0.675	0.297

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment group students who were randomly-selected for the Mindspark scholarship offer but who did not take up the offer have been marked as having 0% attendance, as have all students in the control group. Columns (1) and (2) present OLS value-added models for the full sample, Columns (3) and (4) present IV regressions which instrument attendance with the randomized allocation of a scholarship and include fixed effects for randomization strata, and Columns (5) and (6) present OLS value-added models using only data on the lottery-winners. Scores are scaled here using Item Response theory models and linked across grades and across baseline and endline assessments using common anchor items. Tests in both math and Hindi are standardized to have a mean of zero and standard deviation of one in the baseline.

Table 4: Treatment effect by specific competence assessed

(a) Mathematics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dep var: Proportion of questions answered correctly						
	Arithmetic computation	Word problems - computation	Data interpretation	Fractions and decimals	Geometry and Measurement	Numbers	Pattern recognition
Treatment	0.078***	0.071***	0.044**	0.072***	0.14***	0.15***	0.11***
	(0.016)	(0.016)	(0.020)	(0.020)	(0.026)	(0.023)	(0.029)
Baseline math score	0.13***	0.11***	0.080***	0.090***	0.050***	0.067***	0.094***
	(0.0070)	(0.0095)	(0.013)	(0.011)	(0.014)	(0.012)	(0.013)
Constant	0.66***	0.50***	0.38***	0.33***	0.39***	0.45***	0.36***
	(0.0080)	(0.0077)	(0.0098)	(0.010)	(0.013)	(0.011)	(0.015)
Observations	531	531	531	531	531	531	531
R-squared	0.365	0.227	0.095	0.153	0.092	0.134	0.109

(b) Hindi

	(1)	(2)	(3)	(4)
	Dep var: Proportion of questions answered correctly			
VARIABLES	Sentence completion	Retrieve explicitly stated information	Make straightforward inferences	Interpret and integrate ideas and information
Treatment	0.047*	0.046***	0.064***	0.055***
	(0.024)	(0.016)	(0.022)	(0.016)
Baseline Hindi score	0.13***	0.14***	0.14***	0.064***
	(0.016)	(0.0079)	(0.011)	(0.013)
Constant	0.73***	0.59***	0.52***	0.31***
	(0.012)	(0.0078)	(0.011)	(0.0079)
Observations	533	533	533	533
R-squared	0.186	0.382	0.305	0.132

Note: Robust standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The tables above show the impact of the treatment on specific competences. The dependent variable in each regression is the proportion of questions related to the competence that a student answered correctly. Baseline scores are IRT scores in the relevant subject from the baseline assessment. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. All regressions include randomization strata fixed effects.

Table 5: Heterogeneity in treatment effect by sex, socio-economic status and initial achievement

	(1)	(2)	(3)	(4)	(5)	(6)
		Dep var: Standardized IRT scores (endline)				
VARIABLES	Math	Hindi	Math	Hindi	Math	Hindi
Treatment	0.43***	0.22**	0.36***	0.24***	0.36***	0.22***
	(0.14)	(0.10)	(0.063)	(0.067)	(0.064)	(0.076)
Female	-0.032	0.17				
	(0.15)	(0.16)				
SES index			0.0095	0.088***		
			(0.029)	(0.020)		
Baseline score	0.54***	0.67***	0.54***	0.64***	0.51***	0.67***
	(0.047)	(0.034)	(0.045)	(0.032)	(0.057)	(0.044)
Treatment * Female	-0.082	-0.0037				
	(0.14)	(0.13)				
Treatment * SES index			-0.0011	0.016		
			(0.044)	(0.042)		
Treatment * Baseline score					0.058	-0.0025
					(0.075)	(0.078)
Constant	0.38***	0.021	0.36***	0.15***	0.36***	0.15***
	(0.11)	(0.11)	(0.031)	(0.033)	(0.031)	(0.037)
Observations	529	533	529	533	529	533
R-squared	0.393	0.453	0.393	0.472	0.393	0.451

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. The SES index refers to a wealth index generated using the first factor from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household. Tests in both math and Hindi were designed to cover wide ranges of achievement and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline. All regressions include strata fixed effects.

Table 6: Heterogeneity in treatment effect by within-grade terciles

	(1)	(2)
	Dep var: Standardized IRT scores (endline)	
VARIABLES	Math	Hindi
Bottom Tercile	0.14	-0.11
	(0.091)	(0.10)
Middle Tercile	0.35***	0.11
	(0.073)	(0.078)
Top Tercile	0.57***	0.46***
	(0.086)	(0.079)
Treatment	0.36***	0.34***
	(0.11)	(0.13)
Treatment*Middle Tercile	0.081	-0.21
	(0.15)	(0.17)
Treatment*Top Tercile	-0.040	-0.16
	(0.16)	(0.15)
Baseline test score	0.41***	0.53***
	(0.058)	(0.061)
Observations	529	533
R-squared	0.555	0.516

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. Tests in both math and Hindi were designed to cover wide ranges of achievement and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline.

Table 7: Treatment effect on items linked to grade levels

	(1)	(2)	(3)	(4)
	Dep var: Proportion of questions answered correctly			
	Math		Hindi	
VARIABLES	At or above grade level	Below grade level	At or above grade level	Below grade level
Treatment	0.0023	0.082***	0.069**	0.051***
	(0.039)	(0.012)	(0.024)	(0.013)
Baseline math score	0.044	0.095***		
	(0.025)	(0.0056)		
Baseline Hindi score			0.11***	0.13***
			(0.016)	(0.0065)
Constant	0.31***	0.49***	0.44***	0.58***
	(0.018)	(0.0058)	(0.012)	(0.0065)
Observations	286	505	287	507
R-squared	0.025	0.341	0.206	0.379

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The tables above show the impact of the treatment on questions below or at/above grade levels for individual students. The dependent variable in each regression is the proportion of questions that a student answered correctly. The endline assessments had very few items at higher grade levels and hence we are unable to present estimates of effect on grade-level competences for students in Grades 8 and 9. Baseline scores are IRT scores in the relevant subject from the baseline assessment. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. All regressions include randomization strata fixed effects.

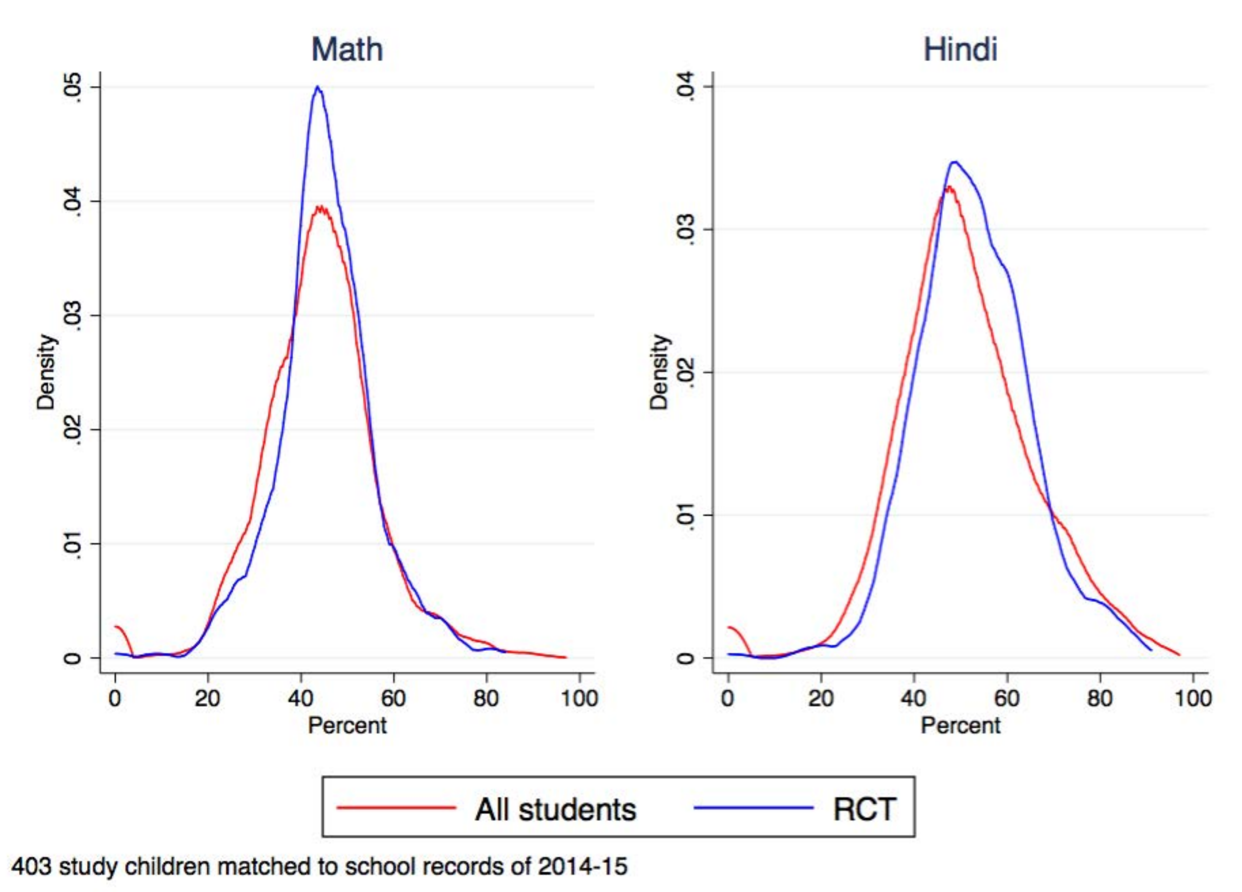
Table 8: Treatment effect on school exams

	(1)	(2)	(3)	(4)	(5)	(6)
			Dep var: Standardized test scores			
VARIABLES	Hindi	Math	Science	Social Sciences	English	Aggregate
Treatment	0.19**	0.058	0.077	0.10	0.080	0.097
	(0.089)	(0.076)	(0.092)	(0.11)	(0.10)	(0.080)
Baseline Hindi score	0.48***		0.28***	0.41***	0.29***	0.33***
	(0.094)		(0.064)	(0.098)	(0.069)	(0.061)
Baseline math score		0.29***	0.10**	0.25***	0.11**	0.16***
		(0.039)	(0.036)	(0.052)	(0.049)	(0.037)
Constant	0.40	0.14	0.88**	0.69	1.11	0.68
	(1.01)	(0.50)	(0.39)	(0.69)	(0.66)	(0.56)
Observations	595	594	593	592	595	595
R-squared	0.188	0.069	0.117	0.173	0.137	0.202

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table shows the effect of receiving the Mindspark voucher on the final school exams, held in March 2016 after the completion of the intervention. The school grades are normalized within school*grade to have a mean of zero and a standard deviation of one in the control group. All regressions include grade and school fixed effects. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. Baseline math and Hindi scores refer to students' scores on the independent assessment administered as part of the study in September 2016.

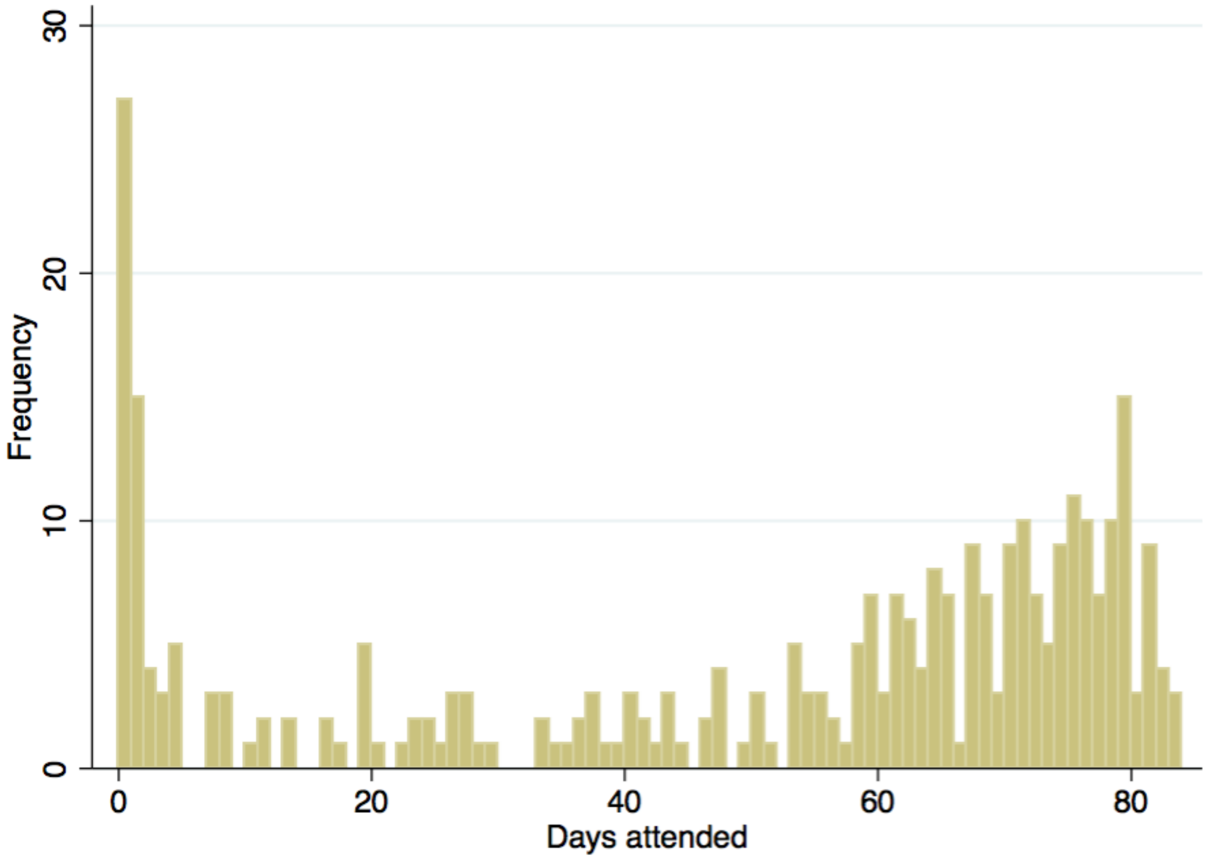
Appendix A Additional figures and tables

Figure A.1: Comparing pre-program achievement of study participants and non-participants



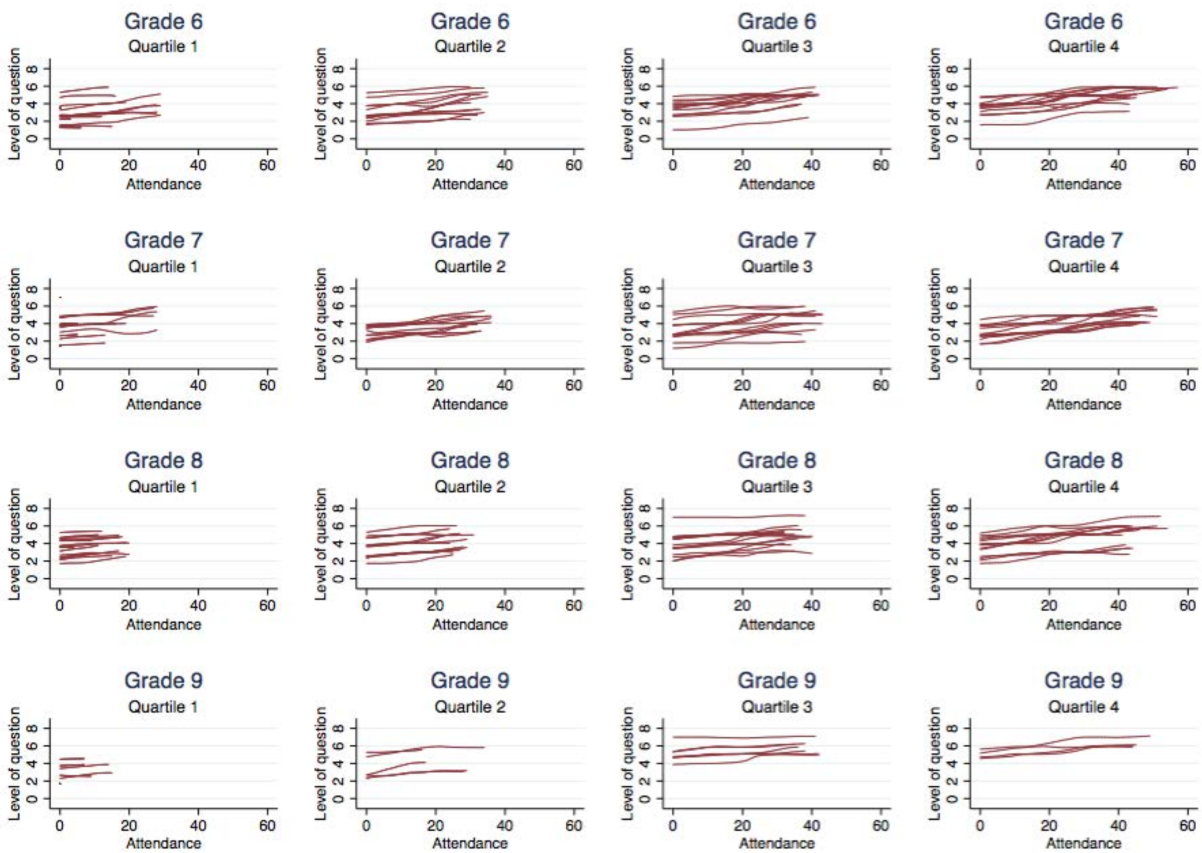
Note: The panels compare the final scores for the 2014-15 school year, i.e. the pre-program academic year, for study participants and non-participants. The study participants seem to be mildly positively selected into the RCT in comparison to their peers but this selection is modest and there is near-complete common support between the two groups in pre-program academic achievement.

Figure A.2: Distribution of take-up among lottery-winners



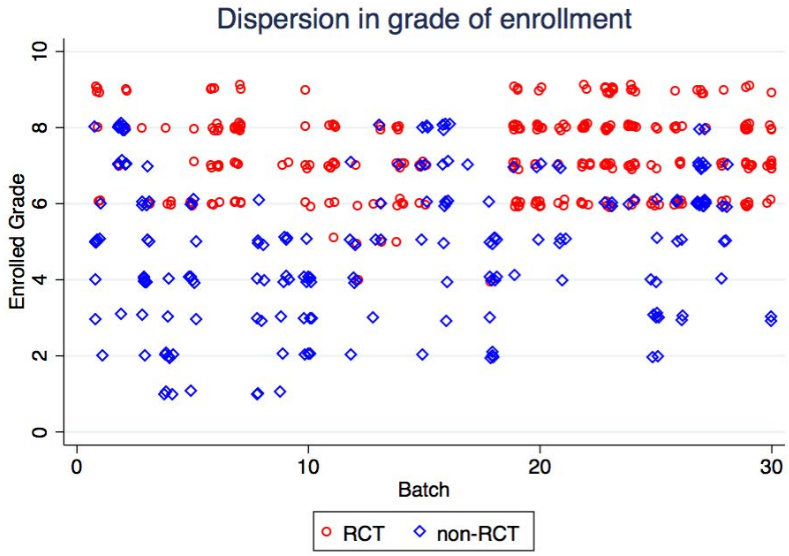
Note: This figure shows the distribution of attendance in the Mindspark centers among the lottery-winners. Over the study period, the Mindspark centers were open for 86 working days.

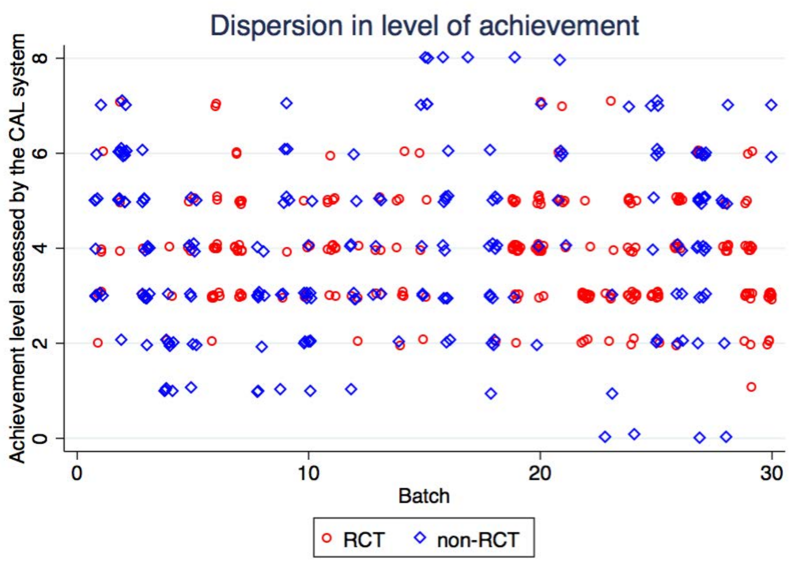
Figure A.3: Learning trajectories of individual students in the treatment group



Note: Each line in the panels above is a local polynomial smoothed plot the grade level of questions administered by the computer adaptive system against Mindspark attendance for an individual child. The panels are organized by the grade of enrolment and the within-grade quartile of attendance in Mindspark.

Figure A.4: Composition of group instruction batches in Mindspark centres





Note: The two panels above show the composition of batches in Mindspark centres by the grade students are enrolled in and by their level of math achievement, as assessed by the Mindspark CAL system. We separately identify students in the treatment group from fee-paying students who were not part of the study but were part of the small group instruction in each batch. Batches are chosen by students based on logistical convenience and hence there is substantial variation in grade levels and student achievement within each batch with little possibility of achievement-based tracking. This confirms that it would not have been possible to customize instruction in the instructor-led small group instruction component of the intervention.

Table A.1: Correlates of attendance

VARIABLES	(1) Attendance (days)	(2)	(3)
Female	3.81	2.51	2.89
	(3.90)	(3.93)	(3.89)
SES index	-3.26***	-3.49***	-3.43***
	(1.04)	(1.07)	(1.06)
Attends private tutoring			-1.95
			(4.41)
Attends Hindi private tutoring			7.27*
			(4.38)
Baseline math score		-1.07	-0.99
		(2.05)	(2.11)
Baseline Hindi score		3.66*	4.17**
		(2.06)	(2.10)
Constant	46.8***	47.7***	45.5***
	(3.39)	(3.42)	(3.79)
Observations	313	310	310
R-squared	0.036	0.045	0.057

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table shows correlates of days attended in the treatment group i.e. lottery-winners who had been offered a Mindspark voucher.

Table A.2: Quadratic dose-response relationship

	(1)	(2)	(3)	(4)
	Full sample		Treatment group	
	Math	Hindi	Math	Hindi
Attendance (days)	0.0056	0.0064	0.0079	0.0064
	(0.0054)	(0.0058)	(0.0073)	(0.0083)
Attendance squared	0.000016	-0.000037	-5.52e-06	-0.000037
	(0.000073)	(0.000078)	(0.000084)	(0.000094)
Baseline math score	0.54***		0.57***	
	(0.039)		(0.062)	
Baseline Hindi score		0.69***		0.68***
		(0.039)		(0.057)
Constant	0.35***	0.15***	0.30**	0.15
	(0.041)	(0.043)	(0.14)	(0.16)
Observations	529	533	261	263
R-squared	0.413	0.468	0.413	0.429

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table models the dose-response relationship between Mindspark attendance and value-added quadratically. Results are estimated using OLS in the full sample and the treatment group only.

Table A.3: Comparing pre-program exam results of study participants and non-participants

	Non-study	RCT	Difference	SE	N(non-study)	N(RCT)
English	45.51	47.06	-1.55**	0.68	4067	409
Hindi	50.67	52.78	-2.12***	0.78	4067	409
Math	43.80	45.28	-1.48**	0.65	4067	409
Science	45.80	46.66	-0.86	0.71	4067	409
Social Science	47.55	49.83	-2.28***	0.64	4067	409

Note: This table presents the mean percentage scores of study participants and non-participants in the 2014-15 school year. Study participants are, on average, positively selected compared to their peers.

Table A.4: Dose-response of Mindspark attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Standardized IRT scores (endline)					
	OLS VA (full sample)		IV models (full sample)		OLS VA (Treatment group)	
VARIABLES	Math	Hindi	Math	Hindi	Math	Hindi
Days of Math instruction	0.018***		0.017***		0.020***	
	(0.0023)		(0.0028)		(0.0047)	
Days of Hindi instruction		0.011***		0.011***		0.0096*
		(0.0026)		(0.0032)		(0.0055)
Baseline score	0.54***	0.69***	0.53***	0.67***	0.56***	0.68***
	(0.039)	(0.039)	(0.036)	(0.037)	(0.061)	(0.056)
Constant	0.35***	0.16***			0.30***	0.18
	(0.040)	(0.042)			(0.12)	(0.13)
Observations	529	533	529	533	261	263
R-squared	0.414	0.469	0.423	0.459	0.414	0.430
Angrist-Pischke F-statistic for weak instrument			1243	1100		
Diff-in-Sargan statistic for exogeneity (p-value)			0.21	0.87		
Extrapolated estimates of 45 days' treatment (SD)	0.81	0.495	0.765	0.495	0.90	0.432

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment group students who were randomly-selected for the Mindspark scholarship offer but who did not take up the offer have been marked as having 0% attendance, as have all students in the control group. Days attended in Math/Hindi are defined as the number of sessions of either CAL or smal group instruction attended in that subject, divided by two. Columns (1) and (2) present OLS value-added models for the full sample, Columns (3) and (4) present IV regressions which instrument attendance with the randomized allocation of a scholarship and include fixed effects for randomization strata, and Columns (5) and (6) present OLS value-added models using only data on the lottery-winners. Scores are scaled here using Item Response theory models and linked across grades and across baseline and endline assessments using common anchor items. Tests in both math and Hindi are standardized to have a mean of zero and standard deviation of one in the baseline.

Table A.5: ITT estimates with within-grade normalized test scores

	(1)	(2)	(3)	(4)
	Dep var: Endline scores			
VARIABLES	Math	Hindi	Math	Hindi
Treatment	0.37***	0.21***	0.36***	0.21***
	(0.067)	(0.067)	(0.068)	(0.073)
Baseline math score	0.56***		0.55***	
	(0.042)		(0.050)	
Baseline Hindi score		0.70***		0.69***
		(0.040)		(0.033)
Constant	0.37***	0.18***	0.37***	0.18***
	(0.046)	(0.046)	(0.033)	(0.036)
Observations	517	521	517	521
R-squared	0.375	0.459	0.376	0.457
Strata fixed effects			Y	Y

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. The SES index refers to a wealth index generated using the first factor from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household. Tests in both math and Hindi were designed to cover wide ranges of ability and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline in each grade.

Table A.6: Treatment effect on take-up of other private tutoring

VARIABLES	(1) Math	(2) Hindi	(3) English	(4) Science	(5) Social Science
Post Sept-2015	0.019*	0.018*	0.026***	0.018**	0.014**
	(0.011)	(0.0096)	(0.0098)	(0.0080)	(0.0071)
Post * Treatment	0.013	-0.010	-0.0039	0.0017	-0.0056
	(0.016)	(0.012)	(0.013)	(0.012)	(0.0086)
Constant	0.21***	0.13***	0.18***	0.14***	0.098***
	(0.0053)	(0.0040)	(0.0044)	(0.0041)	(0.0029)
Observations	3,735	3,735	3,735	3,735	3,735
R-squared	0.009	0.004	0.010	0.007	0.005
Number of students	415	415	415	415	415

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table shows individual fixed-effects estimates of receiving the Mindspark voucher on the take-up in other private tutoring in various subjects. The dependent variable is whether a child was attending extra tutoring in a given month between July 2015 and March 2016 in the particular subject. This was collected using telephonic interviews with the parents of study students. Observations are at the month*child level. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016.

Table A.7: ITT estimates with inverse probability weighting

	(1)	(2)	(3)	(4)
	Dep var: Endline test scores			
VARIABLES	Math	Hindi	Math	Hindi
Treatment	0.37***	0.22***	0.37***	0.23***
	(0.062)	(0.064)	(0.061)	(0.063)
Baseline subject score	0.55***	0.68***	0.54***	0.66***
	(0.039)	(0.040)	(0.037)	(0.038)
Constant	0.36***	0.16***	0.36***	0.16***
	(0.043)	(0.045)	(0.042)	(0.043)
Strata fixed effects			Y	Y
Observations	529	531	529	531
R-squared	0.393	0.455	0.442	0.504

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. Results in this table are weighted by the inverse of the predicted probability of having scores in both math and Hindi in the endline; the probability is predicted using a probit model with baseline subject scores, sex of the child, SES index and dummies for individual Mindspark centres as predictors. Tests in both math and Hindi were designed to cover wide ranges of ability and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline in each grade.

Table A.8: Lee bounds estimates of ITT effects

	(1) Math	(2) Hindi
Lower	0.267	0.21
	(0.119)	(0.127)
Upper	0.445	0.41
	(0.093)	(0.102)
Lower 95% CI	0.071	0.0004
Upper 95% CI	0.599	0.58

Note: Analytic standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ This table presents Lee (2002) bounds on the ITT effects of winning a voucher in both math and Hindi. The bounds are tightened using dummy variables for the Mindspark centres. Tests in both math and Hindi were designed to cover wide ranges of ability and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline in each grade.

Table A.9: ITT estimates, by source of test item

	(1)	(2)	(3)	(4)
	Dep var: Proportion correct in endline			
	Math		Hindi	
VARIABLES	EI items	non-EI items	EI items	non-EI items
Treatment	0.10***	0.071***	0.050***	0.042***
	(0.013)	(0.010)	(0.017)	(0.011)
Baseline score	0.094***	0.096***	0.14***	0.12***
	(0.0096)	(0.0073)	(0.0086)	(0.0058)
Constant	0.46***	0.47***	0.61***	0.48***
	(0.0067)	(0.0049)	(0.0083)	(0.0056)
Observations	531	531	533	533
R-squared	0.228	0.346	0.308	0.403

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark scholarship till March 2016. Tests in both math and Hindi were assembled using items from different international and Indian assessments, some of which were developed by EI. EI developed assessments include the Student Learning Survey, the Quality Education Study and teh Andhra Pradesh Randomized Studies in Education. The dependent variables are defined as the proportion correct on items taken from assessments developed by EI and on other non-EI items. Baseline scores are IRT scores normalized to have a mean of zero and a standard deviation of one.

Appendix B Prior research on hardware and software

Tables B.1 and B.2 offer an overview of experimental and quasi-experimental impact evaluations of interventions providing hardware and software to improve children's learning. The tables only include studies focusing on students in primary and secondary school (not pre-school or higher education) and only report effects in math and language (not on other outcomes assessed in these studies, e.g., familiarity with computers or socio-emotional skills).

B.1 Selecting studies

This does not intend to be a comprehensive review of the literature. Specifically, we have excluded several impact evaluations of programs (mostly, within education) due to major design flaws (e.g., extremely small sample sizes, having no control group, or dropping attritors from the analysis). These flaws are widely documented in meta-analyses of this literature (see, for example, Murphy et al., 2001; Pearson et al., 2005; Waxman et al., 2003).

We implemented additional exclusions for each table. In Table B.1, we excluded DIDs in which identification is questionable and studies evaluating the impact of subsidies for Internet (for example, Goolsbee and Guryan, 2006). In Table B.2, we excluded impact evaluations of software products for subjects other than math and language or designed to address specific learning disabilities (e.g., dyslexia, speech impairment).

B.2 Reporting effects

To report effect sizes, we followed the following procedure: (a) we reported the difference between treatment and control groups adjusted for baseline performance whenever this was available; (b) if this difference was not available, we reported the simple difference between treatment and control groups (without any covariates other than randomization blocks if applicable); and (c) if neither difference was available, we reported the difference between treatment and control groups adjusted for baseline performance and/or any other covariates that the authors included.

In all RCTs, we reported the intent-to-treat (ITT) effect; in all RDDs and IVs, we reported the local average treatment effect (LATE). In all cases, we only reported the magnitude of effect sizes that were statistically significant at the 5% level.[32] Otherwise, we mentioned that a program had "no effect" on the respective subject.[33]

B.3 Categories in each table

In both tables, we documented the study, the impact evaluation method employed by the authors, the sample, the program, the subject for which the software/hardware was designed to

[32]These decisions are non-trivial, as the specifications preferred by the authors of some studies are only significant at the 1% level or only become significant at the 5% level after the inclusion of multiple covariates.

[33]Again, this decision is non-trivial because some of these studies were under-powered to detect small to moderate effects.

target, and its intensity. Additionally, in Table B.1, we documented: (a) whether the hardware provided included pre-installed software; (b) whether the hardware required any participation from the instructor; and (c) whether the hardware was accompanied by training for teachers. In Table B.2, we documented: (a) whether the software was linked to an official curriculum (and if so, how); (b) whether the software was adaptive (i.e., whether it could *dynamically* adjust the difficulty of questions and/or activities based on students' performance); and (c) whether the software provided *differentiated* feedback (i.e., whether students saw different messages depending on the incorrect answer that they selected).

Table B.1: Impact evaluations of hardware

Study	Method	Sample	Program	Subject	Intensity	Software included?	Instructor's role?	Teacher training?	Effect	Cost
Angrist and Lavy (2002)	IV	Grades 4 and 8, 122 Jewish schools in Israel	Tomorrow-98	Math and language (Hebrew)	Target student-computer ratio of 1:10 in each school	Yes, included educational software from a private company	Not specified	Yes, training for teachers to integrate computers into teaching	Grade 4: -0.4 to -0.3σ in math and no effect in language	USD 3,000 per machine, including hardware, software, and setup; at 40 computers per school, USD 120,000 per school
Barrera-Osorio and Linden (2009)	RCT	Grades 3-9, 97 public schools in six school districts, Colombia	Computers for Education	Math and language (Spanish)	15 computers per school	Not specified	Use the computers to support children on basic skills (esp. Spanish)	Yes, 20-month training for teachers, provided by a local university	No effect in language or math	Not specified
Malamud and Pop-Eleches (2011)	RDD	Grades 1-12, in six regions, Romania	Euro 200 Program	Math and language (English and Romanian)	One voucher (worth USD 300) towards the purchase of a computer for use at home	Pre-installed software, but educational software provided separately and not always installed	Not specified	Yes, 530 multimedia lessons on the use of computers for educational purposes for students	-0.44σ in math GPA, -0.56σ in Romanian GPA, and -0.63σ in English	Not specified

Cristia et al. (2012)	RCT	319 schools in eight rural areas, Peru	One Laptop per Child	Math and language (Spanish)	One laptop per student and teacher for use at school and home	Yes, 39 applications including: standard applications, educational games, music editing, programming environments, sound and video recording, encyclopedia; also 200 age-appropriate e-books	Not specified	Yes, 40-hour training aimed at facilitating the use of laptops for pedagogical purposes	No effect in math or language	USD 200 per laptop
Mo et al. (2013)	RCT	Grade 3, 13 migrant schools in Beijing, China	One Laptop per Child	Math and language (Chinese)	One laptop per student for use at home	Yes, three sets of softare: a commercial, game-based math learning program; a similar program for Chinese; a third program developed by the research team	Not specified	No, but one training session with children and their parents	No effect in math or language	Not specified
Beuermann et al. (2015)	RCT	Grade 2, 28 public schools in Lima, Peru	One Laptop per Child	Math and language (Spanish)	Four laptops (one per student) in each class/section for use at school	Yes, 32 applications including: standard applications, educational games, music editing, programming environments, sound and video recording, encyclopedia	Not specified	No, but weekly training sessions during seven weeks for students	No effect in math or language	USD 188 per laptop

Leuven et al. (2007)	RDD	Grade 8, 150 schools in the Netherlands	Not specified	Math and language (Dutch)	Not specified	Not specified	Not specified	Not specified	-0.08 SDs in language and no effect in math	This study estimates the effect of USD 90 per pupil for hardware and software
Machin et al. (2007)	IV	Grade 6, 627 (1999-2001) and 810 (2001-2002) primary and 616 (1999-2000) and 714 (2001-2002) secondary schools in England	Not specified	Math and language (English)	Target student-computer ratio of 1:8 in each primary school and 1:5 in each secondary school	Some schools spent funds for ICT for software	Not specified	Yes, in-service training for teachers and school librarians	2.2 pp. increase in the percentage of children reaching minimally acceptable standards in end-of-year exams	This study estimates the effect of doubling funding for ICT (hardware and software) for a Local Education Authority
Fairlie and Robinson (2013)	RCT	Grades 6-10, 15 middle and high public schools in five school districts in California, United States	Not specified	Math and language (English)	One computer per child for use at home	Yes, Microsoft Windows and Office	No	No	No effect in language or math	Not specified

Table B.2: Impact evaluations of software

Study	Method	Sample	Program	Subject	Intensity	Linked to curriculum?	Dynamically adaptive?	Differentiated feedback?	Effect	Cost
Banerjee et al. (2007)	RCT	Grade 4, 100 municipal schools in Gujarat, India	Year 1: off-the-shelf program developed by Pratham; Year 2: program developed by Media-Pro	Math	120 min./week during or before/after school; 2 children per computer	Gujarati curriculum, focus on basic skills	Yes, question difficulty responds to ability	Not specified	Year 1: 0.35σ on math and no effect in language; Year 2: 0.48σ on math and no effect in language	INR 722 (USD 15.18) per student per year
Linden (2008)	RCT	Grades 2-3, 60 Gyan Shala schools in Gujarat, India	Gyan Shala Computer Assisted Learning (CAL) program	Math	Version 1: 60 min./day during school; Version 2: 60 min./day after school; Both: 2 children per computer (split screen)	Gujarati curriculum, reinforces material taught that day	Not specified	Not specified	Version 1: no effect in math or language; Version 2: no effect in math or language	USD 5 per student per year
Carrillo et al. (2010)	RCT	Grades 3-5, 16 public schools in Guayaquil, Ecuador	Personalized Complementary and Interconnected Learning (APCI) program	Math and language (Spanish)	180 min./week during school	Personalized curriculum based on screening test	No, but questions depend on screening test	Not specified	No effect in math or language	Not specified
Lai et al. (2012)	RCT	Grade 3, 57 public rural schools, Qinghai, China	Not specified	Language (Mandarin)	Two 40-min. mandatory sessions/week during lunch breaks or after school; teams of 2 children	National curriculum, reinforces material taught that week	No, same questions for all students	No, if students had a question, they could discuss it with their teammate, but not the teacher	No effect in language and 0.23σ in math	Not specified
Lai et al. (2013)	RCT	Grades 3 and 5, 72 rural boarding schools, Shaanxi, China	Not specified	Math	Two 40-min. mandatory sessions/week after school; teams of 2 children	National curriculum, reinforces material taught that week	No, same questions for all students	No, if students had a question, they could discuss it with their teammate, but not the teacher	0.12σ in language, across both grades	Not specified

Mo et al. (2014b)	RCT	Grades 3 and 5, 72 rural schools, Shaanxi, China	Not specified	Math	Two 40-min. mandatory sessions/week during computer lessons; teams of 2 children	National curriculum, reinforces material taught that week	No, same questions for all students	No, if students had a question, they could discuss it with their teammate, but not the teacher	0.18σ in math	USD 9439 in total for 1 year
Mo et al. (2014a)	RCT	Grades 3 and 5, 72 rural schools, Shaanxi, China	Not specified	Math	Two 40-min. mandatory sessions/week during computer lessons; teams of 2 children	National curriculum, reinforces material taught that week	No, same questions for all students	No, if students had a question, they could discuss it with their teammate, but not the teacher	Phase 1: no effect in math; Phase 2: 0.3σ in math	USD 9439 in total for 1 year
Lai et al. (2015b)	RCT	Grade 3, 43 migrant schools, Beijing, China	Not specified	Math	Two 40-min. mandatory sessions/week during lunch breaks or after school	National curriculum, reinforces material taught that week	No, same questions for all students	No, if students had a question, they could discuss it with their teammate, but not the teacher	0.15σ in math and no effect in language	USD 7.9-8.8 per child for 6 months
Mo et al. (2016)	RCT	Grade 5, 120 schools, Qinghai, China	Not specified	Language (English)	Version 1: Two 40-min. mandatory sessions/week during regular computer lessons; Version 2: English lessons (also optional during lunch or other breaks); Both: teams of 2 children	National curriculum, reinforces material taught that week	Version 1: No feedback during regular computer lessons; Version 2: feedback from teachers during English lessons	Version 1: if students had a question, they could discuss it with their teammate, but not the teacher; Version 2: feedback from English teacher	Version 1: 0.16σ in language; Version 2: no effect in language	Version 1: RMB 32.09 (USD 5.09) per year; Version 2: RMB 24.42 (USD 3.87) per year

Wise and Olson (1995)	RCT	Grades 2-5, 4 public schools in Boulder, Colorado, United States	Reading with Orthographic and Segmented Speech (ROSS) programs	Language and reading (English)	Both versions: 420 total min., in 30- and 15-min. sessions; teams of 3 children	Not specified	No, but harder problems introduced only once easier problems solved correctly; also in Version 2, teachers explained questions answered incorrectly	No, but students can request help when they do not understand a word	Positive effect on the Lindamond Test of Auditory Conceptualization (LAC), Phoneme Deletion test and Nonword Reading (ESs not reported); no effect on other language and reading domains	Not specified
Morgan and Ritter (2002)	RCT	Grade 9, 4 public schools in Moore Independent School District, Oklahoma, United States	Cognitive Tutor - Algebra I	Math	Not specified	Not specified	Not specified	Not specified	Positive effect (ES not reported) in math	Not specified
Rouse and Krueger (2004)	RCT	Grades 4-6, 4 public schools in urban district in northeast United States	Fast For Word (FFW) programs	Language and reading (English)	90-100 min./day during lessons ("pull-out") or before/after school, 5 days a week, for 6-8 weeks	Not specified	No, but harder problems introduced only once easier problems solved correctly	Not specified	No effect on Reading Edge test, Clinical Evaluation of Language Fundamentals 3rd Edition (CELF-3-RP), Success For All (SFA) test, or State Reading Test	USD 30,000 for a 1-year license for 30 computers, plus USD 100 per site for professional training

Dynarski et al. (2007)	RCT	Grades 4-6, 4 public schools in urban district in northeast United States	Fast For Word (FFW) programs	Language and reading (English)	90-100 min./day during lessons ("pull-out") or before/after school, 5 days a week, for 6-8 weeks	Not specified	No, but harder problems introduced only once easier problems solved correctly	Not specified	No effect on Reading Edge test, Clinical Evaluation of Language Fundamentals 3rd Edition (CELF-3-RP), Success For All (SFA) test, or State Reading Test	USD 30,000 for a 1-year license for 30 computers, plus USD 100 per site for professional training
		Grade 4, 43 public schools in 11 school districts, United States	Leapfrog, Read 180, Academy of Reading, Knowledgebox	Reading (English)	Varies by product, but 70% used them during class time; 25% used them before school, during lunch breaks, or time allotted to other subjects; and 6% of teachers used them during both	Not specified	Not specified, but all four products automatically created individual "learning paths" for each student	Not specified, but all four products provided immediate feedback to students; one provided feedback of mastery; two provided feedback on diagnostics	No effect in reading	USD 18 to USD 184 per student year year (depending on the product)
		Grade 6, 28 public schools in 10 school districts, United States	Larson Pre-Algebra, Achieve Now, iLearn Math	Math	Varies by product, but 76% used them during class time; 11% used them before school, during lunch breaks, or time allotted to other subjects; and 13% of teachers used them during both	Not specified	Not specified, but all three products automatically created individual "learning paths" for each student	Not specified, but all three products provided immediate feedback to students; one provided feedback of mastery; two provided feedback on diagnostics	No effect in math	USD 9 to USD 30 per student year year (depending on the product)

		Algebra I, 23 public schools in 10 school districts, United States	Cognitive Tutor - Algebra I, PLATO Algebra, Larson Algebra	Math	Varies by product, but 94% used them during class time; and 6% of teachers used them during both	Not specified	Not specified, but two products automatically created individual ”learning paths” for each student	Not specified, but all three products provided immediate feedback to students; two provided feedback of mastery; two provided feedback on diagnostics	No effect in math	USD 7 to USD 30 per student year year (depending on the product)
Barrow et al. (2009)	RCT	Grades 8, 10	I Can Learn	Math	Not specified	National Council of Teachers of Mathematics (NCTM) standards and district course objectives	No, but students who do not pass comprehensive tests repeat lessons until they pass them	Not specified	0.17σ in math	30-seat lab costs USD 100,000, with an additional USD 150,000 for pre-algebra, algebra, and classroom management software
Borman et al. (2009)	RCT	Grades 2 and 7, 8 public schools in Baltimore, Maryland, United States	Fast For Word (FFW) Language	Lan-guage and reading (En-glish)	100 min./day, five days a week, for four to eight weeks, during lessons (”pull-out”)	Not specified	No, all children start at the same basic level and advance only after attaining a pre-determined level of proficiency	Not specified	Grade 2: no effect in language or reading; Grade 7: no effect in language or reading	Not specified
Cam-puzano et al. (2009)	RCT	Grade 1, 12 public schools in 2 school districts, United States	Destination Reading - Course 1	Reading (En-glish)	20 min./day, twice a week, during school	Not specified	Not specified	Not specified	No effect in reading	USD 78 per student per year
		Grade 1, 12 public schools in 3 school districts, United States	Headsprout	Reading (En-glish)	30 min./day, three times a week, during school	Not specified	Not specified	Not specified	0.01 SDs in reading (p¿0.05)	USD 146 per student per year

Grade 1, 8 public schools in 3 school districts, United States	PLATO Focus	Reading (English)	15-30 min./day (frequency per week not specified)	Not specified	No, but teachers can choose the order and difficulty level for activities	Not specified	No effect in reading	USD 351 per student per year
Grade 1, 13 public schools in 3 school districts, United States	Waterford Early Reading Program - Levels 1-3	Reading (English)	17-30 min./day, three times a week, during school	Not specified	Not specified	Not specified	No effect in reading	USD 223 per student per year
Grade 4, 15 public schools in 4 school districts, United States	Academy of Reading	Reading (English)	25 min./day, three or more days a week, during school	Not specified	Not specified	Not specified	No effect in reading	USD 217 per student per year
Grade 4, 19 public schools in 4 school districts, United States	LeapTrack	Reading (English)	15 min./day, three to five days a week, during school	Not specified	No, but diagnostic assessments determine "learning path" for each student	Not specified	0.09σ in reading	USD 154 per student per year
Grade 6, 13 public schools in 3 school districts, United States	PLATO Achieve Now - Mathematics Series 3	Math	30 min./day, four days a week, for at least 10 weeks, during school	Not specified	No, but diagnostic assessment determines which activities students should attempt	Not specified	No effect in math	USD 36 per student per year
Grade 6, 13 public schools in 5 school districts, United States	Larson Pre-Algebra	Math	Varies according to the number of topics/weeks in the course, but recommended at least one a week	Not specified	Not specified	Not specified	No effect in math	USD 15 per student per year

		Algebra I, 11 public schools in 4 school districts, United States	Cognitive Tutor - Algebra I	Math	Two days a week (plus textbook three days a week)	Not specified	Not specified	Not specified	No effect in math	USD 69 per student per year
		Algebra I, 12 public schools in 5 school districts, United States	Larson Algebra I	Math	Varies according to the number of topics/weeks in the course, but recommended at least one a week	Not specified	Not specified	Not specified	No effect in math	USD 13 per student per year
Rockoff (2015)	RCT	Grades 6-8, 8 public middle schools in New York, NY, United States	School of One (So1)	Math	Not specified	No, activities sourced from publishers, software providers, and other educational groups	Yes, "learning algorithm" draws on students' performance on each lesson and recommends a "playlist" for each student; at the end of the day, students take a "playlist update"	No, but possibility to get feedback from live reinforcement of prior lessons, live tutoring, small group collaboration, virtual live instruction, and virtual live tutoring	No effect on New York State Math Test or Northwest Evaluation Association (NWEA) test	Not specified

Appendix C Mindspark software

This appendix provides a more detailed description of the working of the Mindspark computer-assisted learning (CAL) software, and specifics of how it was implemented in the after-school Mindspark centers evaluated in our study.

C.1 Computer training

The first time that students log into the Mindspark software, they are presented with an optional routine (taking 10-15 minutes) designed to familiarize them with the user interface and exercises on math or language.

C.2 Diagnostic test

After the familiarization routine, students are presented with diagnostic tests in math and Hindi which are used by the Mindspark platform to algorithmically determine their initial achievement level (at which instruction will be targeted). Tests contain four to five questions per grade level in each subject. All students are shown questions from grade 1 up to their grade level. However, if students answer at least 75% of the questions for their corresponding grade level correctly, they can be shown questions up to two grade levels above their own.[34] If they answer less or exactly 25% of the questions for one grade level above their actual grade, the diagnostic test shows no more questions. Initial achievement levels determined by the Mindspark system on the basis of these tests are only used to customize the first set of content that students are provided. Further customization is based on student performance on these content modules and does not depend on their performance on the initial diagnostic test (which is only used for initial calibration of each student's learning level).

C.3 Math and Hindi content

Mindspark contains a number of activities that are assigned to specific grade levels, based on analyses of state-level curricula. All of the items are developed by EI's education specialists. The Mindspark centers focus on a specific subject per day: there are two days assigned to math, two days assigned to Hindi, one day assigned to English, and a "free" day, in which students can choose a subject.

Math and Hindi items are organized differently. In math, "topics" (e.g., whole number operations) are divided into "teacher topics" (e.g., addition), which are divided into "clusters" (e.g., addition in a number line), which are divided into "student difficulty levels" (SDLs) (e.g., moving from one place to another on the number line), which are in turn divided into questions (e.g., the same exercise with slightly different numbers). The Mindspark software

[34]For example, a grade 4 student will always see questions from grade 1 up to grade 4. However, if he/she answers over 75% of grade 4 questions correctly, he/she will be shown grade 5 questions; and if he/she answers over 75% of grade 5 questions correctly, he/she will be shown grade 6 questions.

currently has 21 topics, 105 teacher topics and 550 clusters. The organization of math content reflects the mostly linear nature of math learning (e.g., you cannot learn multiplication without understanding addition). This is also why students must pass an SDL to move on to the next one, and SDLs always increase in difficulty.

In Hindi, there are two types of questions: "passages" (i.e., reading comprehension questions) and "non-passages" (i.e., questions not linked to any reading). Passage questions are grouped by grades (1 through 8), which are in turn divided into levels (low, medium, or high). Non-passage questions are grouped into "skills" (e.g., grammar), which are divided into "sub-skills" (e.g., nouns), which are in turn divided into questions (e.g., the same exercise with slightly different words). The Mindspark software currently has around 330 passages (i.e., 20 to 50 per grade) linked to nearly 6,000 questions, and for non-passage questions, 13 skills and 50 sub-skills, linked to roughly 8,200 questions. The Hindi content is organized in this way because language learning is not as linear as math (e.g., a student may still read and comprehend part of a text even if he/she does not understand grammar or all the vocabulary words in it). As a result there are no SDLs in Hindi, and content is not necessarily as linear or clearly mapped into grade-level difficulty as in math.

The pedagogical effectiveness of the language-learning content is increased by using videos with same-language subtitling (SLS). The SLS approach relies on a "karaoke" style and promotes language learning by having text on the screen accompany an audio with on-screen highlighting of the syllable on the screen at the same time that it is heard, and has been shown to be highly effective at promoting adult literacy in India (Kothari et al., 2002, 2004). In Mindspark, the SLS approach is implemented by showing students animated stories with Hindi audio alongside subtitling in Hindi to help the student read along and improve phonetic recognition, as well as pronunciation.

C.4 Personalization

C.4.1 Dynamic adaptation to levels of student achievement

In math, the questions within a teacher topic progressively increase in difficulty, based on EI's data analytics and classification by their education specialists. When a child does not pass a learning unit, the learning gap is identified and appropriate remedial action is taken. It could be leading the child through a step-by-step explanation of a concept, a review of the fundamentals of that concept, or simply more questions about the concept.

Figure C.1 provides an illustration of how adaptability works. For example, a child could be assigned to the "decimal comparison test", an exercise in which he/she needs to compare two decimal numbers and indicate which one is greater. If he/she gets most questions in that test correctly, he/she is assigned to the "hidden numbers game", a slightly harder exercise in which he/she also needs to compare two decimal numbers, but needs to do so with as little information as possible (i.e., so that children understand that the digit to the left of the decimal is the most important and those to the right of the decimal are in decreasing order of importance). However, if he/she gets most of the questions in the decimal comparison test incorrectly, he/she is assigned to a number of remedial activities seeking to reinforce fundamental concepts about decimals.

In Hindi, in the first part, students start with passages of low difficulty and progress towards higher-difficulty passages. If a child performs poorly on a passage, he/she is a assigned to a lower-difficulty passage. In the second part, students start with questions of low difficulty in each skill and progress towards higher-difficulty questions. Thus, a student might be seeing low-difficulty questions on a given skill and medium-difficulty questions on another.

C.4.2 Error analysis

Beyond adapting the level of difficulty of the content to that of the student, Mindspark also aims to identify specific sources of conceptual misunderstanding for students who may otherwise be at a similar overall level of learning. Thus, while two student may have the same score on a certain topic (say scoring 60% on fractions), the reasons for their missing the remaining questions may be very different, and this may not be easy for a teacher to identify. A distinctive feature of the Mindspark system is the use of detailed data on student responses to each question to analyze and identify *patterns* of errors in student responses to allow for identifying the precise misunderstanding/misconception that a student may have on a given topic, and to target further content accordingly.

The idea that educators can learn as much (or perhaps more) from analyzing patterns of student errors than from their correct answers has a long tradition in education research (for instance, see citepbuswell1925summary and citepradatz1979error for discussions of the use of "error analysis" in mathematics education). Yet, implementing this idea in practice is highly non-trivial in a typical classroom setting for individual teachers. The power of "big data" in improving the design and delivery of educational content is especially promising in the area of error analysis, as seen in the example below.

Figure C.2 shows three examples of student errors in the hidden numbers game. These patterns of errors were identified by the Mindspark software, and subsequently EI staff interviewed a sample of students who made these errors to understand their underlying misconceptions. In the first example, students get the comparison wrong because they exhibited what EI classifies as "whole number thinking". Specifically, students believed 3.27 was greater than 3.3 because, given that the integer in both cases was the same (i.e., 3), they compared the numbers to the left of the decimal point (i.e., 27 and 3) and concluded (incorrectly) that since 27 is greater than 3, 3.27 was greater than 3.3.

In the second example, the error cannot be because of the reason above (since 27 is greater than 18). In this case, EI diagnosed the nature of the misconception as "reverse order thinking". In this case, students know that the "hundred" place value is greater than the "ten" place value, but also believe that the "hundredth" place value is greater than the "tenth" place value. Therefore, they compared 81 to 27 and concluded (incorrectly) that 3.18 was greater than 3.27.

Finally, the error in the last example cannot be because of either of the two patterns above (since 27 is less than 39, and 7 is less than 9). In this case, EI diagnosed the nature of the misconception as "reciprocal thinking". Specifically, students understand that the component of the number to the right of the decimal is a fraction, but they then proceeded to take the reciprocal of the number to the right of the decimal, the way standard fractions are written. Thus, they were comparing $\frac{1}{27}$ to $\frac{1}{39}$ as opposed to 0.27 to 0.39 and as a result (incorrectly) classified the former as greater.

It is important to note that the fraction of students making each type of error is quite small (5%, 4%, and 3% respectively), which would make it much more difficult for a teacher to detect these patterns in a typical classroom (since the sample of students in a classroom would be small). The comparative advantage of the computer-based system is clearly apparent in a case like this, since it is able to analyze patterns from thousands of students, with each student attempting a large set of such comparisons. This enables both pattern recognition at the aggregate level and diagnosis at the individual student-level as to whether a given student is exhibiting that pattern. Consistent with this approach, Mindspark then targets follow-up content based on the system's classification of the patterns of student errors as seen in Figure C.1 (which also shows how each student would do 30 comparisons in the initial set of exercises to enable a precise diagnosis of misconceptions).

C.5 Feedback

The pedagogical approach favoured within the Mindspark system prioritizes active student engagement at all times. Learning is meant to build upon feedback to students on incorrect questions. Also, most questions are preceded by an example and some "interactives" show step-by-step what students should do.

In math, feedback consists of feedback to wrong answers, through animations or text with voice-over. In Hindi, students receive explanations of difficult words and are shown how to use them in a sentence. The degree of personalization of feedback differs by question: (a) in some questions, there is no feedback to incorrect answers; (b) in others, all students get the same feedback to an incorrect answer; and (c) yet in others, students get different types of feedback depending on the wrong answer they selected.

Algorithms for the appropriate feedback and further instruction that follow a particular pattern of errors are informed by data analyses of student errors, student interviews conducted by EI's education specialists to understand misconceptions, and published research on pedagogy. All decisions of the software in terms of what content to provide after classification of errors are "hard coded" at this point. Mindspark does not currently employ any machine-learning algorithms (although the database offers significant potential for the development of such tools).

In addition to its adaptive nature, the Mindspark software allows the center staff to given an "injection" of items on a given topic if they believe a student needs to review that topic. However, once the student completes this injection, the software reverts to the item being completed when the injection was given and relies on its adaptive nature.

Figure C.1: Mindspark adaptability in math

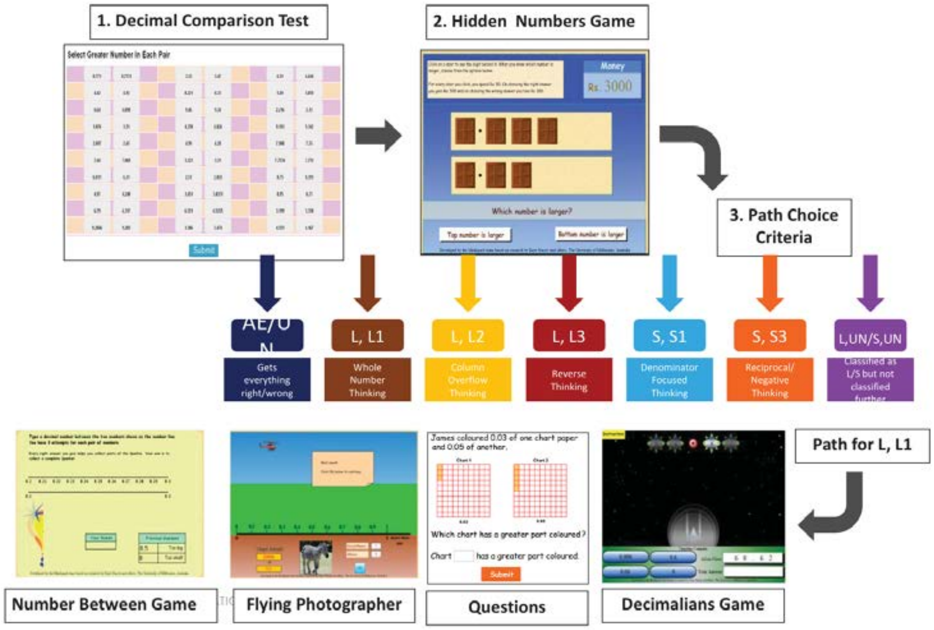


Figure C.2: Student errors in math

Appendix D Test design

D.1 Overview

We measured student achievement, which is the main outcome for our evaluation, using independent assessments in math and Hindi. These tests were administered under the supervision of the research team at both baseline and endline. Here we present details about the test content and development, administration, and scoring.

D.2 Objectives of test design

Our test design was informed by three main objectives. First, was to develop a test which would be informative over a wide range of achievement. Recognizing that students may be much below grade-appropriate levels of achievement, test booklets included items ranging from very basic primary school appropriate competences to harder items which are closer to grade-appropriate standards.

Our secondary objective was to ensure that we were measuring a broad construct of achievement which included both curricular skills and the ability to apply them in simple problems.

Our third, and related, objective was to ensure that the test would be a fair benchmark to judge the actual skill acquisition of students. Reflecting this need, tests were administered using pen-and-paper rather than on computers so that they do not conflate increments in actual achievement with greater familiarity with computers in the treatment group. Further, the items were taken from a wide range of independent assessments detailed below, and selected by the research team without consultation with Education Initiatives, to ensure that the selection of items was not prone to "teaching to the test" in the intervention.

D.3 Test content

We aimed to test a wide range of abilities. The math tests range from simple arithmetic computation to more complex interpretation of data from charts and framed examples as in the PISA assessments. The Hindi assessments included some "easy" items such as matching pictures to words or Cloze items requiring students to complete a sentence by supplying the missing word. Most of the focus of the assessment was on reading comprehension, which was assessed by reading passages of varying difficulty and answering questions that may ask students to either retrieve explicitly stated information or to draw more complex inferences based on what they had read. In keeping with our focus on measuring functional abilities, many of the passages were framed as real-life tasks (e.g. a newspaper article, a health immunization poster, or a school notice) to measure the ability of students to complete standard tasks.

In both subjects, we assembled the tests using publicly available items from a wide range of research assessments. In math, the tests drew upon items from the Trends in Mathematics and Science Study (TIMSS) 4th and 8th grade assessments, OECD's Programme for International Student Assessment (PISA), the Young Lives student assessments administered in four countries including India, the Andhra Pradesh Randomized Studies in Education (APRESt), the India-based Student Learning Survey (SLS) and Quality Education Study (QES); these collectively represent some of the most validated tests in the international and the Indian context.

In Hindi, the tests used items administered by Progress in International Reading Literacy Study (PIRLS) and from Young Lives, SLS and PISA. These items, available in the public domain only in English were translated and adapted into Hindi.

D.4 Test booklets

We developed multiple booklets in both baseline and endline for both subjects. In the baseline assessment, separate booklets were developed for students in grades 4-5, grades 6-7 and grades 8-9. In the endline assessment, given the very low number of grades 4-5 students in our study sample, a single booklet was administered to students in grades 4-7 and a separate booklet for students in grades 8-9. Importantly, there was substantial overlap that was maintained between the booklets for different grades and between the baseline and endline assessments. This overlap was maintained across items of all difficulty levels to allow for robust linking using IRT. Table D.1 presents a break-up of questions by grade level of difficulty in each of the booklets at baseline and endline.

Test booklets were piloted prior to baseline and items were selected based on their ability to discriminate achievement among students in this context. Further, a detailed Item analysis of all items administered in the baseline was carried out prior to the finalization of the endline test to ensure that the subset of items selected for repetition in the endline performed well in terms of discrimination and were distributed across the ability range in our sample. Table D.2 presents the number of common items which were retained across test booklets administered.

D.5 Test scoring

All items administered were multiple-choice questions, responses to which were marked as correct or incorrect dichotomously. The tests were scored using Item Response Theory (IRT) models.

IRT models specify a relationship between a single underlying latent achievement variable ("ability") and the probability of answering a particular test question ("item") correctly. While standard in the international assessments literature for generating comparative test scores, the use of IRT models is much less prevalent in the economics of education literature in developing countries (for notable exceptions, see Das and Zajonc 2010, Andrabi et al 2011, Singh 2015). For a detailed introduction to IRT models, please see Van der Linden and Hambleton (1997) and Das and Zajonc (2010).

The use of IRT models offers important advantages in an application such as ours, especially in comparison to the usual practice of presenting percentage correct scores or normalized raw scores. First, it allows for items to contribute differentially to the underlying ability measure; this is particularly important in tests such as ours where the hardest items are significantly more complex than the easiest items on the test.

Second, it allows us to robustly link all test scores on a common metric, even with only a partially-overlapping set of test questions, using a set of common items between any two assessments as "anchor" items. This is particularly advantageous when setting tests in samples with possibly large differences in mean achievement (but which have substantial common support in achievement) since it allows for customizing tests to the difficulty level of the particular sample but to still express each individual's test score on a single continuous metric. This is particularly important in our application in enabling us to compute business-as-usual value-added in the control group.[35]

Third, IRT models also offer a framework to assess the performance of each test item individually which is advantageous for designing tests that include an appropriate mix of items of varying difficulty but high discrimination.

We used the 3-parameter logistic model to score tests. This model posits the relationship between underlying achievement and the probability of correctly answering a given question as a function of three item characteristics: the difficulty of the item, the discrimination of the item, and the pseudo-guessing parameter. This relationship is given by:

$$P_g(\theta_i) = c_g + \frac{1 - c_g}{1 + exp(-1.7.a_g.(\theta_i - b_g))} \tag{5}$$

where i indexes students and g indexes test questions. θ_i is the student's latent achievement (ability), P is the probability of answering question g correctly, b_g is the difficulty parameter and a_g is the discrimination parameter (slope of the ICC at b). c_g is the pseudo-guessing parameter which takes into account that, with multiple choice questions, even the lowest ability can answer some questions correctly.

Given this parametric relationship between (latent) ability and items characteristics, this relationship can be formulated as a joint maximum likelihood problem which uses the matrix of NxM student responses to estimate $N+3M$ unknown parameters. Test scores were generated

[35] IRT scores are only identified up to a linear transformation. Without explicitly linking baseline and endline scores, the constant term in our value-added regressions (which we interpret as value-added in teh control group) would have conflates the arbitrary linear transformation and value-added in the control group.

using the OpenIRT software for Stata written by Tristan Zajonc. We use maximum likelihood estimates of student achievement in the analysis which are unbiased individual measures of ability (results are similar when using Bayesian expected a posteriori scores instead).

D.6 Empirical distribution of test scores

Figure D.1 presents the percentage correct responses in both math and Hindi for baseline and endline. It shows that the tests offer a well-distributed measure of achievement with few students unable to answer any question or to answer all questions correctly. This confirms that our achievement measures are informative over the full range of student achievement in this setting.

Figure D.2 presents similar graphs for the distribution of IRT test scores. Note that raw percent correct scores in Figure D.1 are not comparable over rounds or across booklets because of the different composition of test questions but the IRT scores used in the analysis are.

D.7 Item fit

The parametric relationship between the underlying ability and item characteristics is assumed, in IRT models, to be invariant across individuals (in the psychometrics literature, referred to as no differential item functioning). An intuitive check for the performance of the IRT model is to assess the empirical fit of the data to the estimated item characteristics.

Figure D.2 plots the estimated Item Characteristic Curve (ICC) for each individual item in math and Hindi endline assessments along with the empirical fit for treatment and control groups separately. The fit of the items is generally quite good and there are no indications of differential item functioning (DIF) between the treatment and control groups. This indicates that estimated treatment effects do not reflect a (spurious) relationship induced by a differential performance of the measurement model in treatment and control groups.

Figure D.1: Distribution of raw percentage correct scores

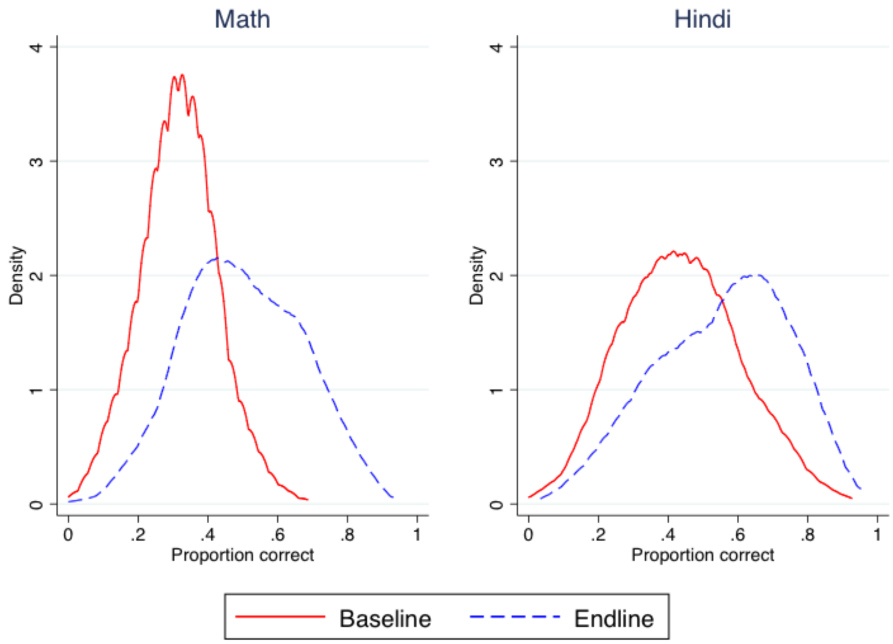


Figure D.2: Distribution of IRT scores, by round and treatment status

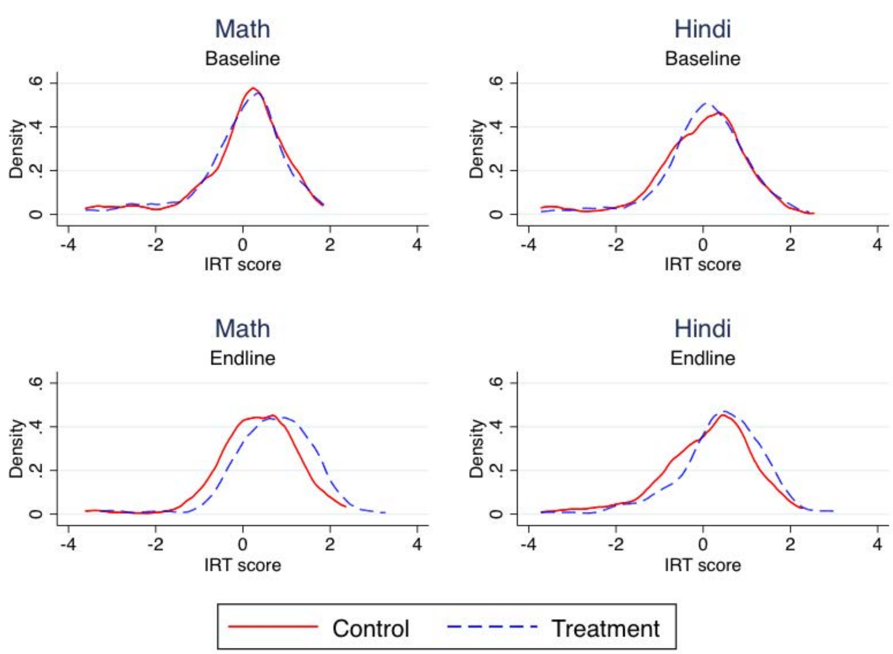
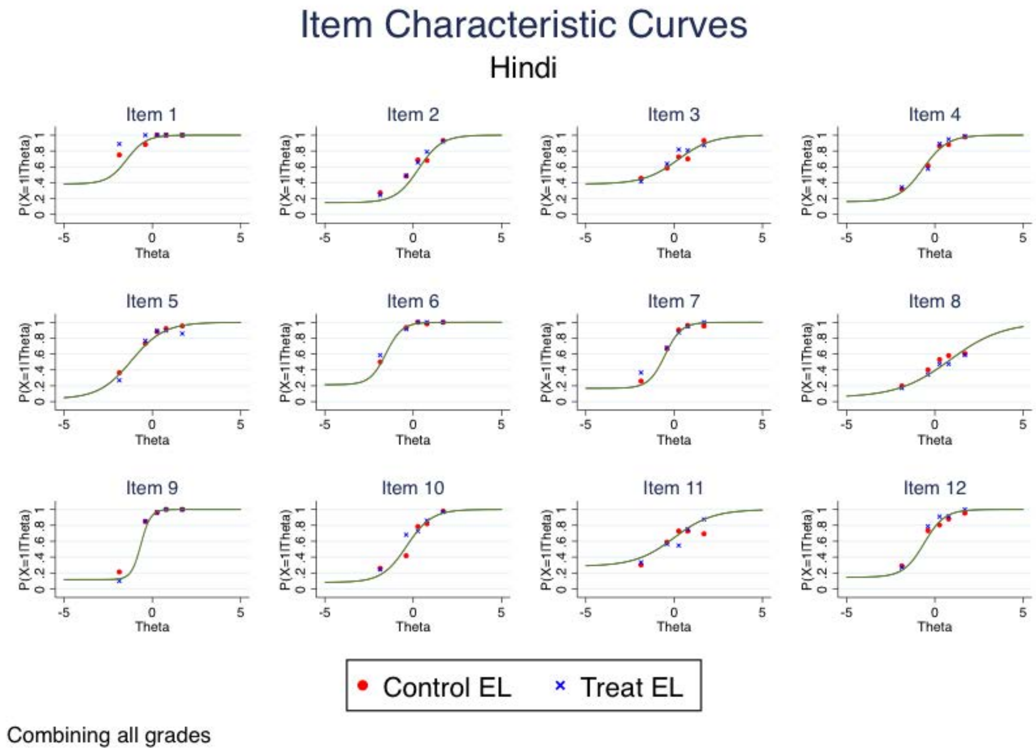
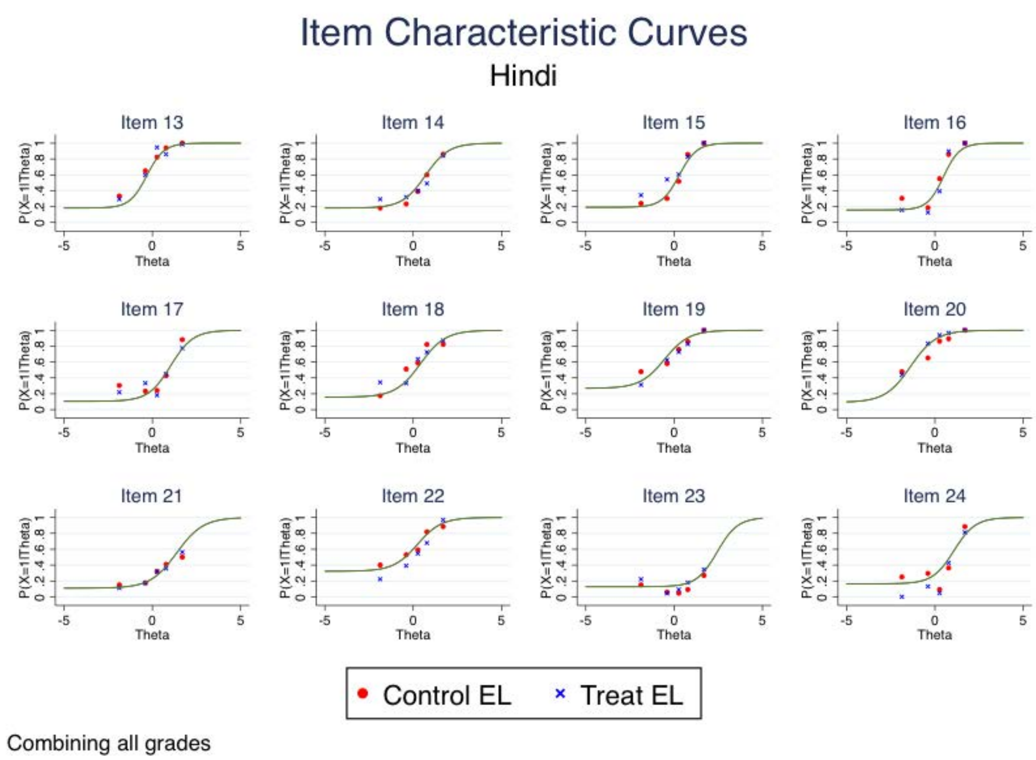


Figure D.3: Item Characteristic Curves: Hindi





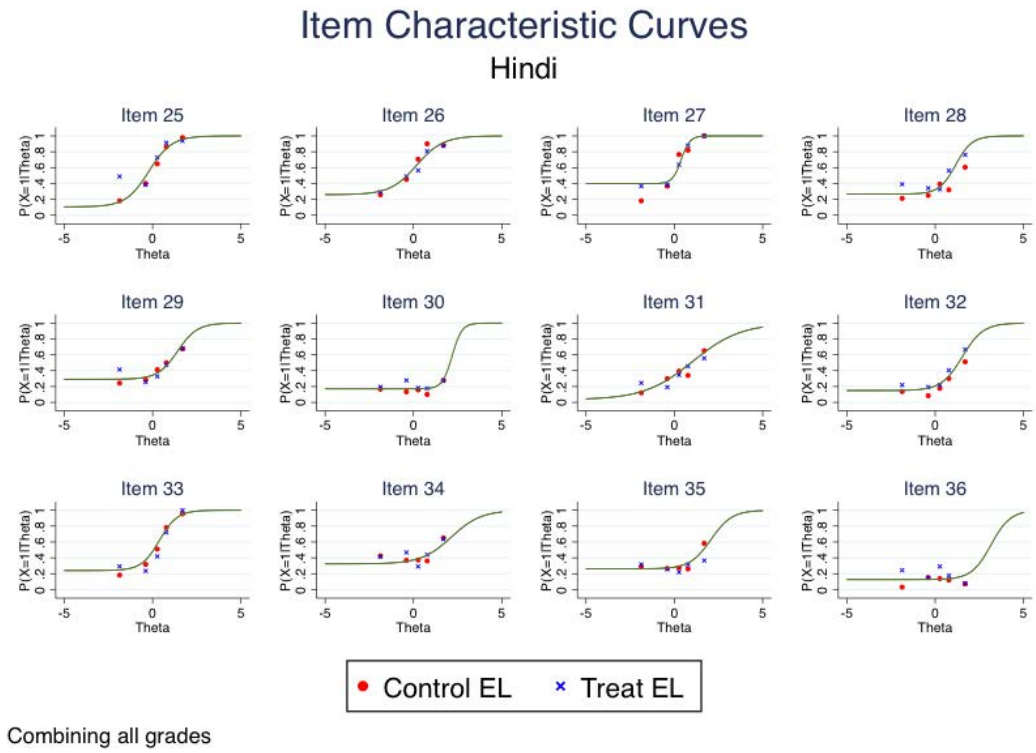
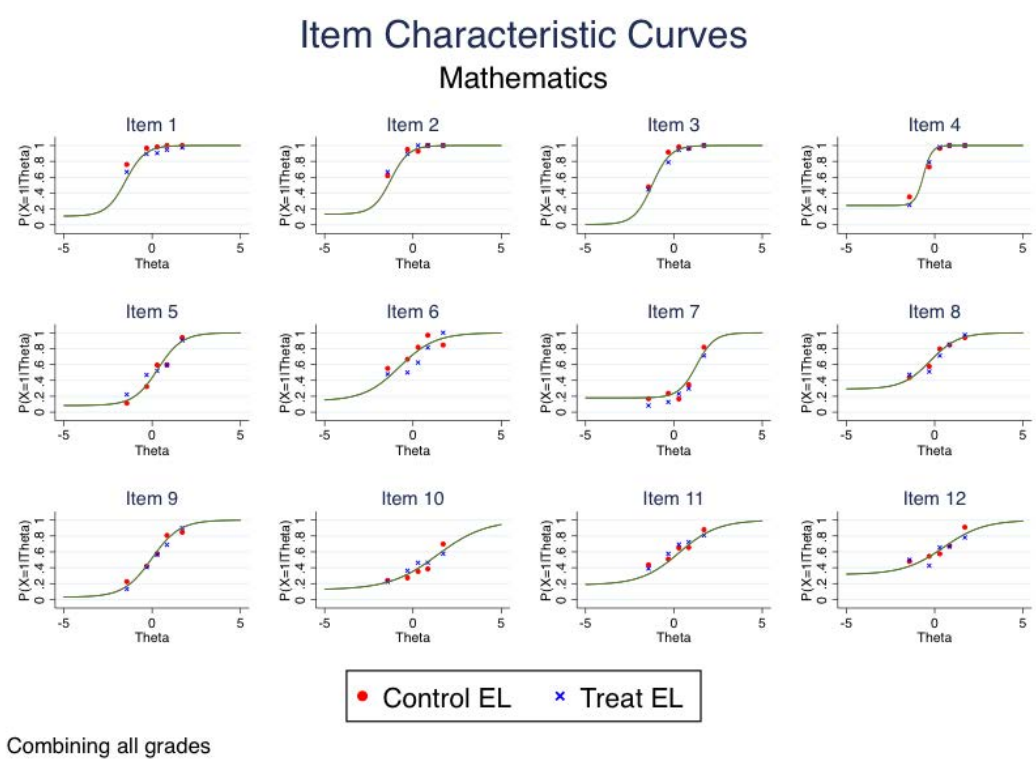
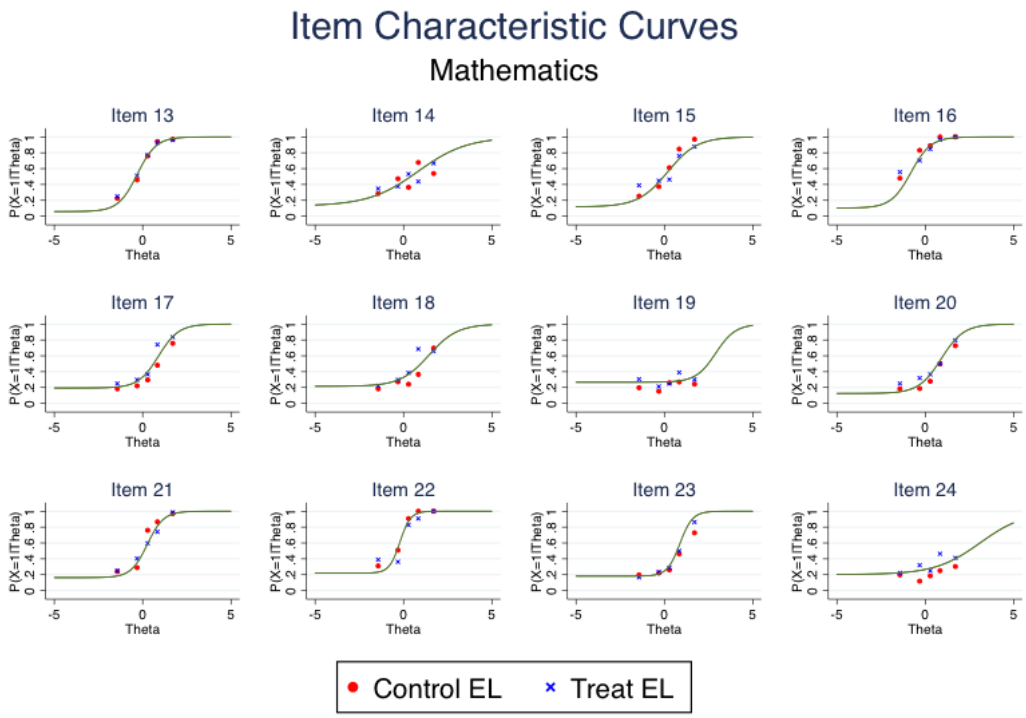
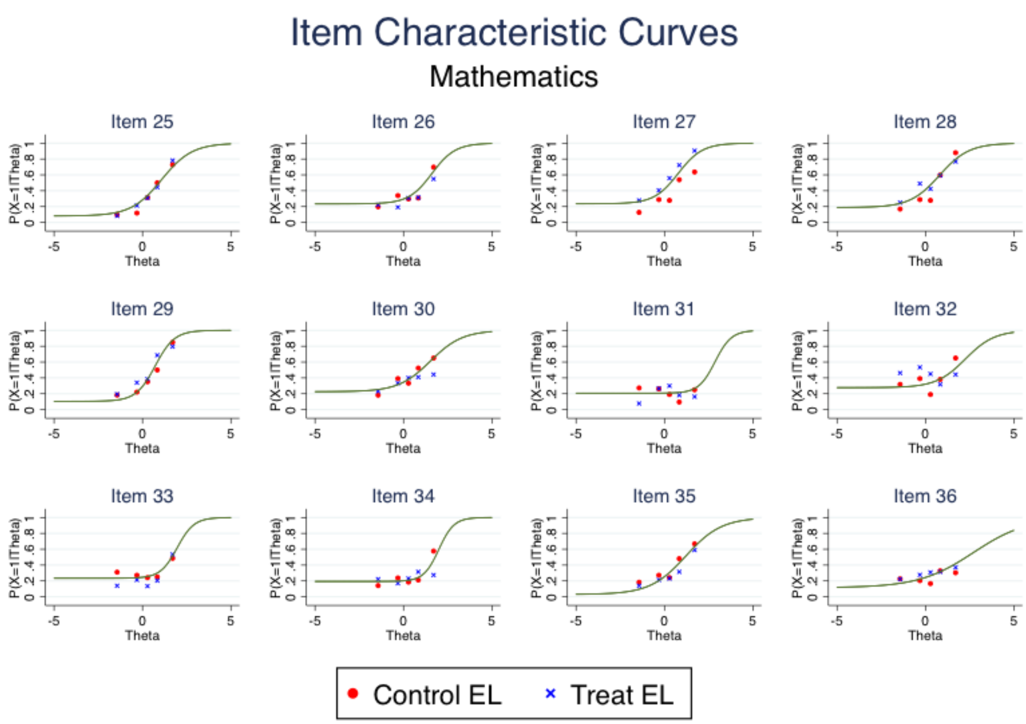


Figure D.4: Item Characteristic Curves: Math





Combining all grades



Combining all grades

Table D.1: Distribution of questions by grade-level difficulty across test booklets

		Booklets				
		Baseline			Endline	
		Math				
		G4-5	G6-7	G8-9	G4-7	G8-9
Number of questions	G2	2	0	0	2	0
at each grade level	G3	14	6	4	6	6
	G4	13	7	4	9	8
	G5	4	10	3	10	10
	G6	1	10	10	5	6
	G7	1	2	11	2	3
	G8	0	0	3	0	2
		Hindi				
		G4-5	G6-7	G8-9	G4-7	G8-9
Number of questions	G2	5	2	1	1	0
at each grade level	G3	3	4	2	1	1
	G4	7	3	3	8	8
	G5	8	7	2	5	6
	G6	0	2	3	11	11
	G7	0	5	9	0	4
	G8	7	7	7	4	0
	G9	0	0	3	0	0

Note: Each cell presents the number of questions by grade-level of content across test booklets. The tests were designed to capture a wide range of student achievement and thus were not restricted to grade-appropriate items only. The grade-level of test questions was established ex-post with the help of a curriculum expert.

Table D.2: Distribution of common questions across test booklets

	Math			
	BL G6-7	BL G8-9	EL G4-7	EL G8-9
BL G4-5	16	10	14	14
BL G6-7		15	10	10
BL G8-9			7	7
EL G4-7				31

	Hindi			
	BL G6-7	BL G8-9	EL G4-7	EL G8-9
BL G4-5	18	10	11	9
BL G6-7		17	13	13
BL G8-9			9	8
EL G4-7				24

Note: Each cell presents the number of questions in common across test booklets. Common items across booklets are used to anchor IRT estimates of student achievement on to a common metric.