Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India*

Karthik Muralidharan† UC San Diego
Abhijeet Singh‡ UCL
Alejandro J. Ganimian§ J-PAL

October 24, 2016

Abstract

Technology-aided instruction has the potential to sharply increase productivity in delivering education, but its promise has yet to be realized. This paper presents experimental evidence on the impact of a technology-aided after-school instruction program on secondary school learning outcomes in urban India. We report five main findings. First, students in this setting are several grade-levels behind their enrolled grade, and this gap grows with every grade. Second, the offer of the program led to large increases in student test scores of 0.36σ in math and 0.22σ in Hindi over a 4.5-month period, which represent a two-fold increase in math and a 2.5 times increase in Hindi test score value-added relative to non-participants. IV estimates suggest that attending the program for 90 days increases math and Hindi test scores by of 0.59σ and 0.36σ respectively. Third, absolute treatment effects are large and similar at all levels of baseline scores, but the relative gain is much greater for academically weaker students because their "business as usual" rate of learning is close to zero. Fourth, we show that the program precisely targets instruction to students' preparation level, thus catering to wide variation within a single grade. Fifth, the program is 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 education by relaxing multiple constraints to effective teaching and learning.

JEL codes: C93, I21, O15.

Keywords: Education technology, computer-aided learning, teaching at the right level, customized learning, post-primary education, India.

*We gratefully acknowledge the funding provided by J-PAL's Post-Primary Education initiative for this project. We thank the staff at Educational Initiatives for their support—especially, Pranav Kothari, Smita Bardhan, Anurima Chatterjee, and Prasid Sreeprakash. We also thank Maya Escueta, Smit Gade, Riddhima Mishra, and Rama Murthy Sripada for research assistance and support. Finally, we thank Abhijit Banerjee, Peter Bergman, Gordon Dahl, Chris Walters as well as seminar participants at CESIfo, IFMR, IGC, NBER Summer Institute, Oxford, UC Berkeley, and UC San Diego for their comments. The usual disclaimers apply.

†Dept. of Economics, University of California, San Diego (UCSD). E-mail: kamurali@ucsd.edu.

‡Dept. of Economics, University College London (UCL). E-mail: abhijeet.singh@ucl.ac.uk.

§Abdul Latif Jameel Poverty Action Lab (J-PAL); E-mail: aganimian@povertyactionlab.org.

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 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 production of human capital lags the rest of the economy, perhaps because the basic technology of classroom-based instruction has changed very little in over 100 years compared to rapid technological progress in other fields (Bosworth 2005).

Thus, it is not surprising that increasing the use of technology in instruction is seen as a leading candidate for improving productivity in education (Gates 2016; Mead 2016). A non-exhaustive list of posited channels of impact include using technology to (a) deliver high-quality content that may circumvent limitations in teachers' own knowledge; (b) deliver engaging (often game-based) interactive content that promotes learning by improving student attention; (c) deliver individually customized learning for students that adjusts materials for both different levels and growth rates of academic progress among students; (d) analyze patterns in student answers to questions to precisely identify areas where students are "stuck" and target instruction appropriately; and (e) sharply reduce the time between students attempting a problem and receiving feedback, which may aid comprehension and understanding.

Yet, despite this theoretical promise, the evidence to date is rather mixed. A recent review of evidence from high-quality studies on the impact of using technology in education globally finds "mixed evidence with a pattern of null results" (Bulman and Fairlie 2016). Thus, while there are many good reasons to be excited about the *potential* for technology-enabled instruction to improve learning outcomes significantly, the evidence suggests that realizing this potential will depend crucially on the details of the specific technology-aided education intervention, and the extent to which it alleviates binding constraints to learning in the status quo. So, a lot more careful research is needed (on both process and impacts) before committing resources to scaling up technology-aided instruction programs - especially in developing country settings with tighter resource constraints.

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 attention to each of the mechanisms listed above. 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, features a database of over 45,000 test questions, and administers

over a million questions to students every day. A key feature of the Mindspark program is the ability to use this data to finely benchmark the baseline learning level of every student and deliver customized content that is targeted at this level, and dynamically adjusts as a function of the rate of progress made by each individual student. Mindspark is platform-agnostic and can be delivered through multiple channels including individual online use, in schools and classrooms, and in after-school programs.

We evaluate the after-school version (delivered through Mindspark centers) in this paper. The program provides a scheduled 90 minutes of instruction, 6 days per week, which is divided into 45 minutes of individual self-driven learning on the Mindspark CAL software and 45 minutes of instructional support from a teaching assistant in groups of 12-15 students.[1] The Mindspark centers aimed to serve students from low and middle-income neighborhoods in Delhi, and charged a modest fee.[2] Our evaluation was carried out in a sample of 619 students recruited for the study from government-run secondary 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) at the beginning and end of the intervention—a gap of about 4.5 months—with assessments linked using item-response theory (IRT) to be comparable on a common scale across the two rounds of testing and across the different grades.

We report five main sets of results. First, we show that students in our sample are several grade-levels behind their grade-appropriate standard, and this gap grows by grade. The average student in grade 6 is an estimated 2.5 years behind curricular levels in Math; by grade 9, this deficit increases to 4.5 years. For the bottom-third of students in the control group, we find that the value-added on our independently-administered tests is close to zero in absolute magnitude 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

[1]The teaching assistant focused on helping students with completing homework and with exam preparation, while the instruction was mostly provided by the Mindspark CAL software. We, therefore, consider our estimates to be a lower bound on the 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).

[2]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 tuition in the vicinity.

not attend for more than a day). Using the lottery as an instrumental variable, we estimate that attending Mindspark 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.

Third, we find that treatment effects do not vary significantly by level of initial achievement, gender or wealth. Thus, consistent with the promise of customized technology-led instruction, the intervention was equally effective in teaching *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 (a) document that there is a very large amount of variation in students' initial preparation for grade-appropriate work, with students enrolled in the same grade typically spanning five to six grades in terms of their readiness, and (b) see that the software targets instruction very precisely to student ability, 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 the ITT estimates) was over 100% greater than the corresponding value-added in the control group and was achieved at substantially less expenditure per student than incurred in the public schooling system. The effectiveness of the Mindspark system is particularly striking when considered in terms of productivity per unit of time. For instance, Muralidharan (2012) finds that providing teachers with individual-level performance-linked bonuses led to student test score gains of 0.54σ and 0.35σ in math and language at the end of 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).[3]

Our first contribution is to the literature in education in developing countries by empirically demonstrating that a key binding constraint in translating education spending into better learning outcomes is the large variation in student preparation and the fact that the level

[3]These cost-effectiveness calculations are done for the full expenditure in money and time on both the computer-based component and the teaching assistant led instructional component. The effect size is likely to be less than full potential of the program since the teacher-led portion was not customized to student levels and had limited integration between the group-instruction component and the computer-based instruction. Further, our cost effectiveness estimates are also likely too conservative for assessing the full potential of the program since the cost-per-child of the program declines sharply with scale. These issues are discussed in greater detail in Section 5

and pace of instruction envisaged by the curriculum and textbooks may be too high for most students. Previous studies had found suggestive evidence of this variation (Banerjee and Duflo 2012; Pritchett and Beatty 2015), but ours is the first to document it empirically.

Second, we contribute to the literature on computer-aided learning (CAL), where the evidence to date appears mixed (Bulman and Fairlie 2016). Nevertheless, some clear patterns are starting to emerge. Hardware-focused interventions that provide computers at home or at school seem to have very little 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). Interventions that focus on improved pedagogy and allowing 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.[4] However, interventions that use technology to also personalize instruction seem to deliver substantial gains. For instance Banerjee et al. (2007) test the impact of a CAL program that allowed some personalization on basic math skills, and find test-score gains of 0.47σ in 2 years.[5] Our results are consistent with this and demonstrate the potential for well-designed technology-aided instruction to deliver large test-score gains in both math and language in a short period of time. For a detailed summary of this literature in developing and developed countries, please see Appendix B.

Third, our results speak more broadly to the potential for technology to accelerate the development process by enabling developing countries to leapfrog constraints to human development. For instance, Deaton (2013) documents that life expectancy in developing countries is much higher than historical levels in Western societies at comparable stages of development, and suggests that these are likely due to improvements in medical technology (such as vaccinations and antibiotics) that are available now. Glewwe and Muralidharan (2016) document that this is also true for enrollment in formal schooling, but not yet true for learning outcomes. Our results point to the possibility that technology-aided instructional solutions could eventually contribute to a similarly positive result for learning outcomes.[6] One approach that has been successful in addressing the variation in student preparation has been that of "Teaching at the Right Level", which uses volunteers to group primary school students by their level of preparation and teach them basic skills (Banerjee et al. 2016, 2007). However, it is not clear if this approach can be extended to secondary grades where the content

[4]See, for example, Barrow et al. (2009); Carrillo et al. (2010); Lai et al. (2015, 2013, 2012); Linden (2008); Mo et al. (2014); Rouse and Krueger (2004).

[5]The customization was limited because two students shared a computer, but the program provided math games whose level of difficulty adjusted to the pace of the pair of students working together. To our knowledge, the only CAL program that most closely resembles the features of the Mindspark software is the one evaluated by Rockoff (2015).

[6]Examples for 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 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).

is more advanced and complex, and the level of variation in student preparation is much higher (exacerbated by "social promotion" policies in many countries). These conditions make the effective delivery of *any* curriculum challenging and our results suggest that technology-aided instruction may be especially effective in such settings with large intra-class variation.

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 Intervention

Mindspark is a computer-assisted learning (CAL) software designed to provide personalized instruction to children in primary and secondary school. It was developed by Educational Initiatives, a leading Indian private assessment firm established in 2001 with considerable experience designing, administering, and analyzing student assessments at the national and state levels. Mindspark has been deployed through: (a) stand-alone, after-school centers in low-income areas; (b) a dedicated part of the school day in government and private schools; and (c) a self-paced online platform. The software is platform-agnostic and can be delivered through computers, tablets, and smartphones, both online and offline.

We evaluated a version of Mindspark delivered through three stand-alone centers in Delhi. Children attend these centers after school (if they go to school in the morning) or before school (if they go to school in the afternoon). This version of the program provides students with 45 minutes of the CAL software and 45 minutes of instructor-led small group instruction (SGI). [7] Children sign up for the program by selecting a "batch" (i.e., 90-minute slot), which includes about 12 to 15 students. Typically, parents pay INR 200 (USD 3) per month to send their children to the program.

2.1.1 Computer-assisted learning

In the 45 minutes allotted to the CAL software, each child is assigned to a computer with a software 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.

One of the distinctive features of this software is that it *adaptive* (i.e., the difficulty of the activities presented to each child are based on that child's performance). The software is able

[7] The intensity of the program was designed to be comparable to private extra tuition, which is common in India. According to the 2012 India Human Development Survey, 43% of 11-17 year olds attended paid extra tuition outside of school.

to customize the level of difficulty at a very granular level, drawing on more than 45,000 items developed by EI. This adaptation occurs both at the beginning of the program, and then with every subsequent activity the children complete. On their first session, children complete a diagnostic test that assesses their initial learning level and determines the difficulty level of the first set of activities that they will see (i.e., children who perform poorly in the test will see easier items, and those who perform well will see harder items). Once the children begin their first activity, the software dynamically adjusts the difficulty of each subsequent activity based on their performance up until each day.

Another important feature of the software is that it provides *differentiated feedback* based on analysis of patterns of students' errors. EI has a dedicated team of item developers who specialize in math, Hindi, or English. EI staff regularly analyze which questions children answer incorrectly, and when an incorrect answer is chosen frequently, they interview a sample of students to understand the source of their misconception. Then, based on this information, EI adjusts the message that the software displays when children select incorrect answers. Some questions display the same message for all incorrect answers in a given question, while others display different messages for each incorrect option.

The software is also *interactive* to promote students' conceptual understanding of the material (instead of rote memorization or mechanical application of procedures). It is not a series of test questions, but rather a set of games, videos, and activities from which children learn through explanations and feedback. This facilitates children's engagement with the material and allows them to progress at their own pace.

Appendix C provides further details about the CAL software, including more information on the diagnostic test, grade-wise content, and error diagnostics.

2.1.2 Small group instruction

In the 45 minutes allotted to SGI, an instructor teaches all students in a batch. This was not an original component of the program, but it was added in response to parental demand to offer children help with their homework and exam preparation. According to EI, instructors often develop a personal relationship with the children, which helps to ensure that they attend the centers regularly.

Instructors are locally-hired. They are selected 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 an extensive library of guiding documents and videos. They are paid the minimum legal wage for a full time person (about USD 100-200 per month), and the center manager is paid slightly more (USD 250-500 per month). In addition to teaching during the

SGI, instructors also supervise children during the time allotted to the CAL component and have access to analytics on children's performance on the CAL software.

The content taught during the SGI component covers core concepts of relative broad relevance for all children. Instructors do not cater to individual learning levels because children typically select their batch based on their school schedule, so each batch includes children from different grades and levels. When possible, EI staff in the Mindspark centers attempt to reassign children across batches to reduce heterogeneity in performance levels, but the logistical feasibility of such reassignment is low since batch preferences are usually dictated by other commitments of the students (such as school hours, any other tuition or domestic responsibilities).

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 from state secondary schools managed by the Government of the National Capital Territory of Delhi (GoNCTD) that were close to the Mindspark centers. School visits and student recruitment was carried out in September 2015. Prior authorization to approach schools was obtained from the Directorate of Education and the recruitment of study participants was conducted in five schools closest to the Mindspark centers in which school principals agreed to the recruitment of participants.[8] Of these five schools, three were girls-only schools and the other two were boys-only secondary schools. 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. Of the potential population of 6,460 students enrolled in grades 4-9 in these schools, 766 showed up for the demonstration sessions.

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 on a scheduled day of testing 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 the duration of most of the school year. Students who

[8]The Delhi Government provided a list of 15 schools that could potentially be targeted for study recruitment. Of these, seven were discarded as they were too far from the Mindspark centers or because the contact information provided was incorrect. The research team contacted the other eight schools, of which five agreed for their students to participate in the study.

were not chosen by lottery would be provided the scholarship after February 2016, conditional on also participating in an endline assessment in February 2016.

Of the 766 students who attended the demonstration sessions, 695 showed up for the baseline assessments but only 619 completed all sections and were thus included in the study. About 97.5% of the study participants were enrolled in grades 6-9.[9] Given that the study sample of 619 students is a self-selected sample of the total eligible population of students in these grades in the targeted schools, we may be concerned at how representative they are of the broader population of students in these grades. In Figure A.1, we present the test score distribution of study participants and non-study participants in the final exams in the preceding school year (2014-15) which were matched from administrative school records. While study participants have slightly better final 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 also 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.[10] Characteristics of the treatment and control group students is presented in Table 1 along with p-values from two-tailed t-tests of equality of means. The treatment and control groups do not differ significantly in any observable dimension. Of the 314 students offered a scholarship for the Mindspark program, over 80% attended the program for at least 7 days.[11]

[Insert Table 1 here.]

Of the 619 students who participated in the baseline test, 533 also attended the endline test (270 control students and 263 treatment students), i.e. yielding a follow-up rate of about 86%, which is somewhat higher in the control group but not statistically significantly different between treatment and control students at the 5% level of significance. Student characteristics at baseline, including test scores, are not statistically significantly different even in the sub-sample of students who later participated in the endline tests.

[9] 15 students in the sample were reported as enrolled in grades 4 and 5 in total with 589 students enrolled in grades 6-9. The enrolled grade was not reported for 15 students.

[10] 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.

[11] There is, however, wide variation in the number of days attended which will be looked at when discussing the main program effects in Section 4.

3 Data

3.1 Student achievement

The primary outcome measure for this study is student achievement. Student achievement was 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 independent monitoring by J-PAL staff to ensure integrity of the assessments.

The tests were designed independently by the research team and intended to capture a wide range of ability in anticipation of wide variance in the achievement of students. 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 advanced international assessments.

Test questions were taken from independent assessments previously administered by high-quality research projects in India (such as Young Lives and the Andhra Pradesh Randomized Studies in Education) and internationally-validated assessments (such as the Trends in Mathematics and Science Study, the Program for International Student Assessment and the Progress in International Reading Literacy Study). 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. Details of the test design and scoring are provided in Appendix D. The assessments performed well in capturing a wide range of ability with very few students being subject to ceiling or floor effects.

3.2 Mindspark CAL system data

The Mindspark CAL system collects detailed logs of all interactions that an individual child has with the software platform. This includes, for example, the daily attendance of each student, the estimated student ability level 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 details of interaction such as time taken to answer or keystrokes to measure engagement with content.

These data are available for the treatment group for the duration of the intervention. We shall be using them in three ways: to describe the distribution of grade deficits in each grade at baseline; to demonstrate the personalization of instruction at the core of the Mindspark

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

system; and to characterize the evolution of student ability in the treatment group over the period of the treatment.

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 school test scores).

3.4 Student characteristics

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 tuitions. At endline, we additionally captured information about a few variables such as their attendance of paid private tuition which may potentially have changed in response to participation in the study or being allocated treatment.

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 as well as some basic information about their opinion of the program.

4 Results

4.1 Business-as-usual academic progress

Our first results characterize the context in which our intervention takes place and describe the progress of academic achievement under business-as-usual settings. Figure 1 shows, for the treatment group, the full joint distribution of grade currently enrolled in and the actual level of student attainment as assessed by the Mindspark CAL system at the start of treatment.[13]

[Insert Figure 1 here.]

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

[13]The Mindspark CAL system benchmarks the academic preparation levels of students to grade levels based on a common assessment taken by newly-enrolled students at the beginning of the intervention. All students take the same test which begins at grade 1 level, benchmarked using direct links to official curricula, and presents sets of items of increasing difficulty. These items go at least up to the current grade the student is enrolled in and may, depending on student performance, go up to two grades beyond their current level of enrollment. This initial benchmarking is then used mainly to customize instruction level in the CAL system and to provide a diagnostic to Mindspark center staff. The benchmarking is done based on a pre-determined formula which weights proportion correct at items at different grade levels.

about 2.5 grades behind in math and about half a grade behind in Hindi.[14] Second, although 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 much flatter than the line of equality between curricular standards and actual achievement levels. This indicates that typical student progress is considerably behind curriculum-expected norms. As a result, the gap between actual student proficiency and the curriculum is greater in later grades with students being nearly 4.5 grades behind in math by grade 9 and 2.5 grades behind in Hindi. By grade 8 or 9, not even the top students in our sample seem to be at grade-appropriate competence in either subject. And third, the figure also 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. This wide level of variation is difficult for any individual teacher to manage in the classroom but may potentially be resolved by personalized instruction targeted at the individual child's learning level.

We next present, as a descriptive measure of student progress, 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. Specifically, we estimate a regression of the form (without a constant term):

$$Y_{is2} = \alpha.terc_{is1} + \gamma.Y_{is1} + \epsilon_{i2} \tag{1}$$

where Y_{ist} is student i's test score on our independent asessment in subject s at period t, $terc_{is1}$ is a vector of indicator variables for the within-grade terciles of baseline achievement in the given subject s and ϵ is the error term. [15]

Coefficients from the vector α, which may be interpreted as the value-added in each tercile, are presented in Figure 2. Students at different parts of the distribution make very different progress — initially better-achieving students also have much 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.

[14] The math portion of the software was developed much earlier and therefore has much more precise linking to curricular expectations than Hindi, which was developed more recently and for which the benchmarking is perhaps less robust. The allocation of test questions to grade levels is also much more robust in math than language (where competencies are less well-delineated across grades). Thus, although most patterns across grades are similar across the two subjects, the computer system's assessment on grade level competence of children is likely to be more reliable for math. For both subjects, we verified that baseline test scores on our independent tests increase significantly with each successive assessed grade level of achievement (as per the CAL program). This indicates that the benchmarking does present similar variation as our independent tests.

[15] Test scores are normalized to have a mean of zero and a standard deviation of one in the baseline in the pooled sample. Standard errors, in this regression and throughout this paper, are heteroskedasticity-corrected robust (Huber-White) standard errors.

[Insert Figure 2 here.]

These two figures confirm that being left behind the curriculum is the typical experience for students, rather than an aberration. This problem grows progressively more severe in later grades, as curriculum difficulty significantly outpaces the growth in student achievement, and a substantial minority of lower-performing students make no academic progress at all.

To the best of our knowledge, we are the first to present direct evidence on these patterns taken together in developing countries. However, they do agree with much indirect evidence on low achievement in Indian schools and the relatively slow progress in growth of achievement in repeated cross-sections (see e.g. Pritchett (2013)). Thus, although our sample is not representative and we use these results mostly to set the context for the intervention, we see little reason to think that the "business-as-usual" performance in this sample is particularly atypical.

4.2 Main program effects

4.2.1 Intent-to-treat estimates

Figure 3 presents the mean test scores in the baseline and endline assessments in both subjects for the lottery-winners and the lottery-losers. While test scores improve between baseline and endline for both groups, endline test scores are significantly and substantially higher for the treatment group indicating much greater academic progress.

[Insert Figure 3 here.]

Our core specification for examining intent-to-treat (ITT) treatment effects is as follows:

$$Y_{is2} = \alpha + \beta_1 . Treatment_i + \gamma . Y_{is1} + \phi_i + \epsilon_{it} \tag{2}$$

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.

ITT effects estimated from Specification (2) are large - at 0.37σ in math and 0.23σ in Hindi - and statistically significant at the 1% level (Cols. 1-2, Table 2).

[Insert Table 2 here.]

In Cols. 3 and 4, we omit strata fixed effects from the regression, noting that the constant term in this case provides an estimate of the value-added in the control group over the course of the

treatment.[16] 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

The ITT results in Table 2 are estimated with an average attendance of about 50 days among lottery-winners (out of a maximum possible attendance of 86 days).[17] These are thus likely to be an underestimate of the program effects under full compliance.

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

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

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.[18]

We first estimate this using OLS i.e. as a lagged value-added (VA) model. Results from this specification show a strong and significant relationship between the the number of days attended and the value-added over the study period in both subjects (Table 3, Cols. 1-2). Our results also indicate that variation in attendance is able to account for the full extent of the ITT treatment effects; the constant term in the OLS value-added regressions, is near-identical to our estimates of value-added in the control group in Table 2.

[Insert Table 3 here.]

However, recognizing that attendance may be endogenous to expected gains from the program, we further instrument attendance by the random offer of a scholarship.[19] The coefficient

[16] Interpreting the constant in this manner is made possible because the baseline and endline tests are linked to a common metric using Item Response Theory. Such an interpretation would not be tenable if scores were normalized within grade/period as is common practice. Our treatment effects, however, are of very similar magnitudes when scores are normalized using a within-grade normalization instead.

[17] 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 with lower baseline achievement in Hindi appear to have greater attendance but the implied magnitudes are small. The full distribution of attendance among lottery-winners is presented in Figure A.2.

[18] Lottery-losers were not allowed to enrol in Mindspark over the duration of the study but were guaranteed the scholarship upon conclusion of the study.

[19] Note that the random offer of a scholarship in our case moves some lottery-winners from zero to positive attendance i.e. the identification of the IV is coming from the extensive margin of attendance.

on days attended is practically unchanged and we cannot reject the null that attendance is conditionally exogenous when controlling for baseline achievement (Cols. 3-4).[20] The coefficient is also stable when estimated as an OLS VA model using only data on the treatment group (Cols. 5-6). [21]

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, for those whose treatment status is affected by the instrument" (Angrist and Imbens 1995). Without further assumptions, we cannot extrapolate what the effect of varying treatment intensity would have been. In particular, extrapolation requires assumptions about (a) the nature of heterogeneity in treatment effects and (b) the functional form of the relationship between days attended and the treatment effect.

The stability of the coefficients on attendance across all specifications is suggestive that, in this particular instance, the average causal response corresponds closely to the marginal effect of an additional day as estimated in the value-added model. We will further be documenting the absence of evidence of treatment effect heterogeneity along observable dimensions in Section 4.4. In this context, however, a constant treatment effects assumption appears justifiable.

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 to attendance at the higher range of our sample, we cannot reject a linear dose-response relationship.[22]

[Insert Figure 4 here.]

Assuming constant treatment effects and a linear dose-response function, both of which appear reasonable in this application, we can provide suggestive magnitudes of the treatment effect under alternative intensities of treatment.[23] Under these assumptions, our results suggest that

[20]This is evident from the high p-values of the Difference-in-Sargan test statistic of the null that attendance is, in fact, exogenous. The test is conducted under the maintained assumption that the IV is valid, which in our case is justified through randomized assignment of the voucher.

[21]OLS VA models may be thought of as a dynamic treatment effects estimator which rely on lagged achievement for ignorability (Guarino, Reckase and Wooldrige, 2015). The close correspondence here between the VA and IV results adds to much recent evidence that VA estimates typically agree closely with experimental and quasi-experimental estimates (see, for example, Andrabi et al. 2011; Angrist et al. 2015; Chetty et al. 2014; Deming 2014; Deming et al. 2014; Kane et al. 2013; Singh 2015, 2016).

[22]We test this explicitly in Table A.3 where we regress the endline test score on attendance, the square of attendance and the baseline test score. The quadratic term is statistically indistinguishable across all specifications (although this may partially reflect low statistical power).

[23]Note that these assumptions do not underlie our identification of the program effects. They are made here only to facilitate extrapolation of estimates to varying dosage of treatment.

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.[24]

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 with students). In Appendix Table A.4, we present analogous IV and value-added specifications which only take into account the time spent by students on either computer-aided or small-group instruction in the particular subject (Math or Hindi) to their learning gains in that subject; using these estimates, 90 days of instructional time in the two subjects, split equally, would lead to treatment gains of 0.765 SD in math and 0.495 SD in Hindi.

4.3 What competences do students learn?

Our tests were assembled to include items of widely differing levels of difficulty and posed different tasks for students. It is additionally useful to understand the specific competences that the intervention has improved. We classified each question in our endline tests by the domain it measured. Using this classification, we present the treatment effects expressed as proportion correct in each type and domain of question. These results are presented in Table 4.

[Insert Table 4 here.]

The intervention led to significant increases across all domains of test questions. 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).

4.4 Heterogeneity

Table 5 presents an investigation of whether treatment effects vary for boys and girls, for richer students vs. poorer students and for better-prepared vs. weaker students. We find no significant evidence of such heterogeneity.

[Insert Table 5 here.]

[24] A school year has an average of about 220 working days. We extrapolate results to half a school year, rather than a full school year, because the implied number of days is close to the maximum number of days attended observed in our sample and thus we are not extrapolating too far out of the range of our data.

In Figure 5, we present a non-parametric representation of the ITT effect plotting kernel-weighted local polynomial smoothed lines, which relate absolute endline test scores to percentiles in the baseline achievement distribution, separately for the treatment and control groups. In both math and Hindi, the trajectory of achievement is shifted upwards for the 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.

[Insert Figure 5 here.]

Effects may still be heterogeneous by the relative position in the within-grade achievement distribution.[25] We investigate this by regressing endline achievement on the baseline test score; indicator variables for the treatment and for the tercile 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.

[Insert Table 6 here.]

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.[26]

4.5 Personalization

The computer-based instruction in Mindspark combines effects from multiple channels: uniformly high quality content, personalization of instruction to students' individual academic preparation and pace of learning, shorter feedback loops with prompt remediation of errors

[25] Two students enrolled in different grades may have the same absolute achievement but occupy different ranks in their respective within-grade distributions. Such heterogeneity could exist for multiple reasons: the curriculum could be targeted at the top end of the within-class achievement distribution; teachers could focus effort on better-performing students; or perhaps doing better/worse might lead to changes in students' self-efficacy and thereby effort. For the effect of ordinal position in the within-grade distribution (see, for example, Weinhardt and Murphy 2016).

[26] This follows naturally from the observation that for weaker students the same absolute effect is being divided by a smaller denominator.

and misconceptions, and a possibly more engaging format that increases student engagement. 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 examine the claim that the Mindspark system does precisely target instructional material at an individual student's level. Direct evidence for this is presented in Figure 6. We present, 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 math in a single day.[27] 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.

[Insert Figure 6 here.]

Second, we examine the claim that the Mindspark CAL system updates its estimate of student achievement levels in real time and constantly updates instruction accordingly. Thus it accommodates variation, not just in the incoming levels of student preparedness (as indicated in the previous figure), but also in their pace of learning and individual trends in student achievement. Evidence of this is presented in Figure 7 where we present non-parametric plots of the difficulty level of the math items presented to students over the course of the intervention.[28] 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 ability. 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.

[27] In both math and Hindi, this is student achievement from a single day which is near the beginning of the intervention but on a day that all students would have completed their initial assessment and, crucially for ensuring an adequate sample size, a day when Mindspark computer-aided instruction in the relevant subject was scheduled in all three centers.

[28] We study this issue in the data on the math questions only. This is for two reasons. First, the dynamic adaptation in math is more finely developed, over a much longer period of time, than in Hindi. Second, while dynamic adaptation in math is focused at moving students to a harder question in the same competence, conditional on answering initial question(s) correctly, in Hindi the software is focused on making sure that at each grade level a student has a mastery of all basic competences before being presented questions at the next grade level.

[Insert Figure 7 here.]

We can however study this pattern at even greater granularity. As a final piece of evidence, which serves also to highlight the richness of the big data on student achievement available to us, we separately plot the learning trajectory of each individual child from the treatment group who attended Mindspark. We present this in separate panels for each grade and by the quartile of attendance in Mindspark in Figure 8. There are three key patterns in these very disaggregated graphs: (a) while there is wide variation in initial ability, we see a general increase in the grade level of the questions across individual trajectories; (b) the increase in assessed ability levels by attendance increases over the entire range of attendance in this sample and (c) most importantly, not only can the program deal with wide variation in the level of initial preparedness, it can also accommodate extensive variation in the pace of learning. The pattern that the slopes of the lines are slightly different across the groups and over time reflects that Mindspark was constantly revising the pace of the instruction to individual needs.

[Insert Figure 8 here.]

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.

4.6 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. This is made more likely with precise personalization of content to student levels, as students are possibly faced with little 'grade-appropriate' instruction. Thus, a relevant question is whether the intervention increased student performance on school tests. [29]

We first use the CAL software data to look directly at the grade level of the material presented by Mindspark to students in the treatment group in both math and Hindi (see Figure 9). The

[29] This is important both because school tests may be high-stakes and with long-term implications (such as matriculation exams at the end of high school) and because, in a context where many students are first-generation learners, school tests may be the metric most salient to parents to judge the academic performance of their children.This matters because parental investments are likely to respond to their judgments about how the child is performing. See, for example, Dizon-Ross (2014) who documents that parents in Malawi invest resources (enrollment and secondary school scholarships) based on their assessment of children's academic achievement, choosing to concentrate these resources on children they believe are performing well in school.

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.

[Insert Figure 9 here.]

Next, we explore the treatment effects expressed at the proportion of test questions answered correctly at grade level and at below-grade level.[30] This is presented in Table 7 for both math and Hindi. As can be seen, the patterns differ across subjects. In math, 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, we find that the treatment effect is significant in all regressions and of meaningful magnitudes.

[Insert Table 7 here.]

Table 8 presents the treatment effect of being offered a voucher on scores on the school exams held in March 2016.[31] Mirroring the results on grade-level items on our own test, 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.

[Insert Table 8 here.]

4.7 Other extra tuition

A final issue to explore is whether the program crowded out extra tuition among the treatment group. This is relevant because private after-school tuition is common in this setting and in our sample. It is plausible that the offer of a voucher for Mindspark may have crowded out paid extra tuition among the treatment group. It also possible that, upon losing the lottery, control groups increase their uptake of private tuition. Both these effects would lead our treatment effects to be lower than implied production function parameters.

[30] Our tests were not designed to be linked to the curriculum and were, rather, designed to capture a wide range of ability. Ex-post, with the help of a curricular expert, we classified each item on our tests as belonging uniquely to a particular grade-level.

[31] March is the end of the academic year in India when students sit end-of-year exams. In Delhi, these exams are taken on a standardized common question paper 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.

We use data from phone surveys of the parents of the study children to investigate this issue. Specifically, we collected information on whether the student attended extra tuition (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 estimate the following regression where each observation is a month/child observation:

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

where T_{ism} is an indicator variable for whether child i attended extra tuition in subject s in month m, $Treatment$ is an indicator variable witha value of one for all lottery-winners and post is an indicator variable for being in a time period after September 2016. λ_{is} is a set of individual fixed effects.

Results from this specification are presented in Table 9. As is evident, we find no evidence of substitution away from extra tuition in this sample.

[Insert Table 9 here.]

4.8 Interpreting a composite treatment effect

The intervention, as administered, is bundled by design and we cannot isolate the individual effects of group instruction and computer-based instruction. However, a major goal of the Mindspark intervention is to deal with substantial variation in academic preparation between children, even within the same grade, and the computer software is key to achieving this. We have shown the customization in the program both to the levels and pace of learning of individual students. It is unlikely that this could be achieved even by motivated teachers, when faced by a classroom with students of very mixed ability. Thus, although we cannot provide causal decompositions of the treatment effect, we think it likely that a substantial portion results directly from the CAL component.

This is particularly the case when we consider the structure of the group teaching. As noted in the program description, students in the group were mixed in age, ability and the grade they are enrolled in. The instruction focused mostly on homework support or the revision of primary school level foundational skills but was not optimized for individual students or, except in an informal sense, to fully utilize the insights from the CAL data. According to EI, 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.

We thus interpret our results as reflecting the evaluation of a "blended learning" model with a particularly well-developed CAL component but with the instructor-led component not fully optimized for teaching. Thus, our results are best interpreted as lower-bound estimates of the full potential of this composite class of interventions. In assessing the cost effectiveness and productivity in terms of time below, which are the relevant parameters for policy, we will always account for the full expenditure in time and money on the program.

5 Cost-effectiveness

Since the Mindspark centers program was offered after school, a natural comparison is with after-school private tuition, which is commonplace in India and in many other developing countries. In a contemporaneous study to ours, Berry and Mukherji (2016) conduct an experimental evaluation of private tuition 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 the similarities, this intervention differs from ours in two significant respects, both of which are central to the posited theory-of-change behind Mindspark - instruction is delivered at grade-level curriculum and not customized to the ability of the child and, secondly, the instruction is delivered in person by a tutor in groups of up to 20 students (similar to the small-group instruction component of the Mindspark centers). Both these features are typical of existing private tuition 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, in the same context, with students at the same grade levels, with a highly motivated NGO implementer 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.

A second comparison is with the productivity of government-run schools (from where the study subjects were recruited). The 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.[32]

[32]These are the estimated value-added in the control group in Table 2 and also include the effects of home inputs and private tuition, and is therefore an upper-bound of learning gains in the public schooling system.

The full costs of the Mindspark program as delivered were about INR 1000 per student (approximately USD 15) per month. These include the costs of infrastructure, hardware and staffing as well as development costs for the Mindspark program and are much higher than the likely steady-state costs because the centers operated at a small scale and substantially below capacity during the study period. Using the 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. 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.

More generally, the promise of platforms like Mindspark for improving education in developing countries at scale comes from the fact that the majority of the costs above reflect fixed costs of product development, and so the *marginal* cost of extending the intervention is much lower. Per-pupil costs decline very sharply with scale: If implemented in government schools, the costs of the program (including hardware costs but excluding rent and utilities) reduce to about USD 25 per child *per year* at even a very modest a scale of 100 schools; at a scale of 1000 schools, these reduce to about USD 9.5 annually per-child. Considering the costs of the software and associated technical support alone, the per-pupil cost at scale is expected to be below USD 2 annually, which is much lower than the USD 150 annual cost (over 10 months) during our pilot. Further, our IV estimates suggest that the gains from attending Mindspark regularly would also be higher than the ITT estimates used in the calculations above.

Of course, our results may not be replicated if Mindspark is implemented within schools at scale, and so our experiment should be treated as an efficacy trial that should be followed up by further evaluations at a larger scale and over a longer duration. But these results are very timely because while there is much interest in policy circles in India and other developing in using technology in education, most of the funds are being deployed to purchase hardware with very little focus on how this technology should be deployed for effective pedagogy.[33] In a context where hundreds of millions of dollars may be spent on the purchase of computer hardware and the creation of IT labs anyway (which evidence suggests is unlikely to improve student learning by itself), the marginal cost of deploying the Mindspark software would be particularly low. Such a deployment would offer a natural opportunity to test impacts at a larger scale.

[33]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.

6 Conclusions

In this paper, we have presented an experimental evaluation of a technology-led supplementary instruction program targeted at improving learning outcomes in post-primary grades. We show 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 s*tudents 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 academic preparation of students and in handling wide variation in the academic levels of students enrolled 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. Personalized instruction makes it possible to accommodate large deficits in initial student preparation and wide variation within a single grade. The high quality of content, combined with effective delivery and interface, circumvents issues of the constricted availability of effective and motivated instructors. Efficient algorithms for error correction, administered in real-time, allow 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.[34] It may be possible to optimize teacher-led instruction more closely on the extensive information on the performance of students, individually and in a group, than is currently the practice in Mindspark centers. 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.

[34] Average attendance rate varies widely across Indian states and across schools, including across private and government schools. The problem of low absolute productivity is, however, near-universal including in most private schools (Muralidharan and Sundararaman 2015; Singh 2015). Thus, at least in some states/sectors, attendance rates might well be much better in schools than in the Mindspark centers intervention.

Finally, the detailed and continuous measures of effort input by the students can be used directly to incentivize students, with potentially large gains in student achievement.[35]

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).[36] Identifying the most effective modes of delivery for the program is likely to be a useful avenue of future enquiry.[37] Our present study is best regarded as an efficacy trial documenting proof-of-concept rather than an endorsement for wholesale adoption. Thus it is important that any attempts to expand the use of such education technology be rigorously evaluated.[38]

Our results have broader relevance for current issues in education in developing countries. First, while there has been a contentious 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.[39] Our results offer insight in this area. 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.[40] Over-ambitious curricula, the neglect of weak performance in most of the distribution, and the

[35]Direct evidence that this may be possible is provided by Hirshleifer (2015) who uses data from a (different) computer-aided instruction intervention to incentivize 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.

[36]For instance, see Linden (2008) and the discussion in Taylor (2015).

[37]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' ability 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.

[38]For broader recent discussions of why this is important, please see, for example, Deaton and Cartwright (2016) and Muralidharan et al. (2016).

[39]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.

[40]For a stark 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

focus on a small well-performing minority are, plausibly, all symptoms of misaligned priorities. 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.

Finally, our results raise 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 extra tuition in India at all levels of education. But the Mindspark centers themselves did not generate substantial 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 consistent under-subscription. Is it, perhaps, that parents are not well-informed of the efficacy of the intervention?[41] Or are they not willing to pay for instruction that improves learning outcomes but may not improve, at this late stage, their performance in high-stakes matriculation exams and (possibly) their chances for securing coveted formal sector employment?[42] While encouraging the reallocation of public expenditure away from less productive uses like above-inflation pay increases (see, for example de Ree et al. 2015) might be welfare-improving, there may also be large unrealized gains in prompting more optimal allocation of household resources, even in poorer settings.

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

[41]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).

[42]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, Jensen 2010, 2012; Munshi and Rosenzweig 2006). 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 an experiment, see Dupas and Miguel (2016) on health and Berry and Mukherji (2016) in education.

References

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

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

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

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

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

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

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

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics 122*(3), 1235–1264.

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

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

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

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

Beuermann, D. W., J. Cristia, S. Cueto, O. Malamud, and Y. Cruz-Aguayo (2015). One Laptop per Child at home: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics 7*(2), 53–80.

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

Bosworth, B. (2005). The Internet and the university. 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. Fairlie (2016). Technology and education: Computers, software and the internet. In E. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, pp. 239–280. Elsevier.

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

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

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

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

Cristia, J., P. Ibarrarán, S. Cueto, A. Santiago, and E. Severín (2012). Technology and child development: Evidence from the One Laptop per Child program. (IDB Working Paper No. IDB-WP-304). Washington, DC: Inter-American Development Bank.

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

de Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2015). Double for nothing? experimental evidence on the impact of an unconditional teacher salary increase on student performance in indonesia. (NBER Working Paper No. 21806). Cambridge, MA: National Bureau of Economic Research (NBER).

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

Deaton, A. and N. Cartwright (2016). Understanding and misunderstanding randomized controlled trials. (NBER Working Paper No. 22595). Cambridge, MA: National Bureau of Economic Research (NBER).

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

Deming, D. J., J. S. Hastings, T. J. Kane, and D. O. Staiger (2014). School choice, school quality, and postsecondary attainment. *American Economic Review 104*(3), 991–1013.

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

Dupas, P. and E. Miguel (2016). Impacts and determinants of health levels in low-income countries. (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 (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort. *Unpublished manuscript*. Washington, DC: Mathematica Policy Research.

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

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

Gates (2016). Finding what works: Results from the LEAP innovations pilot network. Seattle, WA: Bill and Melinda Gates Foundation.

Glewwe, P. and K. Muralidharan (2016). Improving school education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In E. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, pp. 653–744. Elsevier.

Goolsbee, A. and J. Guryan (2006). The impact of internet subsidies in public schools. *The Review of Economics and Statistics 88*(2), 336–347.

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

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

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

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

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

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Seattle, WA: Bill & Melinda Gates Foundation.

Koppensteiner, M. F. (2014). Automatic grade promotion and student performance: Evidence from Brazil. *Journal of Development Economics 107*, 277–290.

Lai, F., R. Luo, L. Zhang, and S. Huang, Xinzhe Rozelle (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education 47*, 34–48.

Lai, F., R. Luo, L. Zhang, X. Huang, and S. Rozelle (2015). Does computer-assisted learning improve learning outcomes? evidence from a randomized experiment in migrant schools in beijing. *Economics of Education Review 47*, 34–48.

Lai, F., L. Zhang, X. Hu, Q. Qu, Y. Shi, Y. Qiao, M. Boswell, and S. Rozelle (2013). Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi. *Journal of Development Effectiveness 52*(2), 208–231.

Lai, F., L. Zhang, Q. Qu, X. Hu, Y. Shi, M. Boswell, and S. Rozelle (2012). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai, China. (REAP Working Paper No. 237). Rural Education Action Program (REAP). Stanford, CA.

Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics 89*(4), 721–736.

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

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

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

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

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

Mo, D., Y. Bai, M. Boswell, and S. Rozelle (2016). Evaluating the effectiveness of computers as tutors in china.

Mo, D., J. Swinnen, L. Zhang, H. Yi, Q. Qu, M. Boswell, and S. Rozelle (2013). Can one-to-one computing narrow the digital divide and the educational gap in China? The case of Beijing migrant schools. *World development 46*, 14–29.

Mo, D., L. Zhang, R. Luo, Q. Qu, W. Huang, J. Wang, Y. Qiao, M. Boswell, and S. Rozelle (2014). Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi. *Journal of Development Effectiveness 6*, 300–323.

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

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

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

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

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

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

Muralidharan, K. and V. Sundararaman (2015). The aggregate effect of school choice: Evidence from a two-stage experiment in India. *The Quarterly Journal of Economics 130*(3), 1011–1066.

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

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

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

Pritchett, L. and A. Beatty (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development 40*, 276–288.

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

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

Singh, A. (2015). Private school effects in urban and rural india: Panel estimates at primary and secondary school ages. *Journal of Development Economics 113*, 16–32.

Singh, A. (2016). Learning more with every year: School year productivity and international learning divergence. *Unpublished manuscript*. London, UK: University College London.

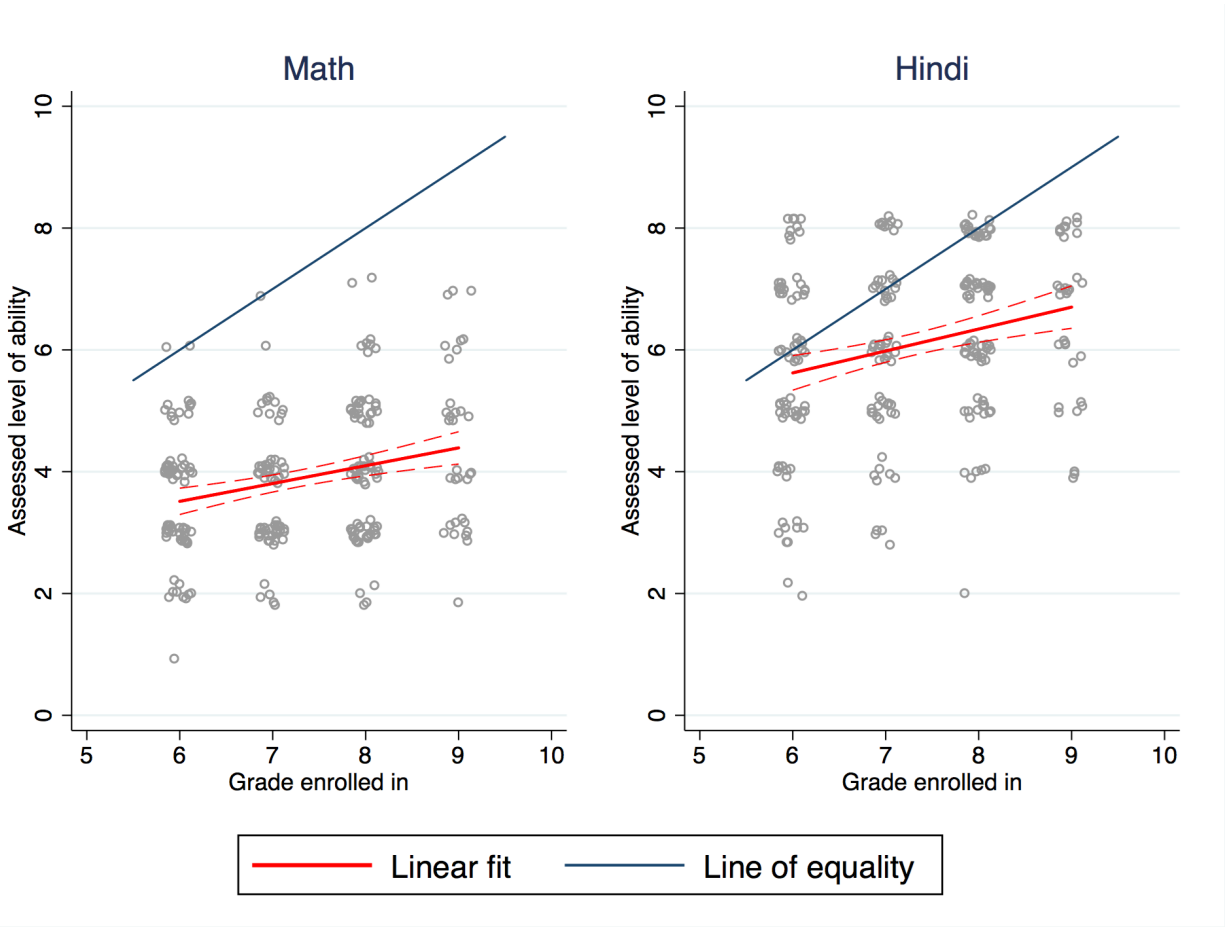
Taylor, E. S. (2015). New technology and teacher productivity. *Unpublished manuscript*. Cambridge, MA: Harvard Graduate School of Education.

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

Weinhardt, F. and R. Murphy (2016). Top of the class: The importance of ordinal rank. (CESifo Working Paper No. 4815). Munich, Germany: CESIfo.

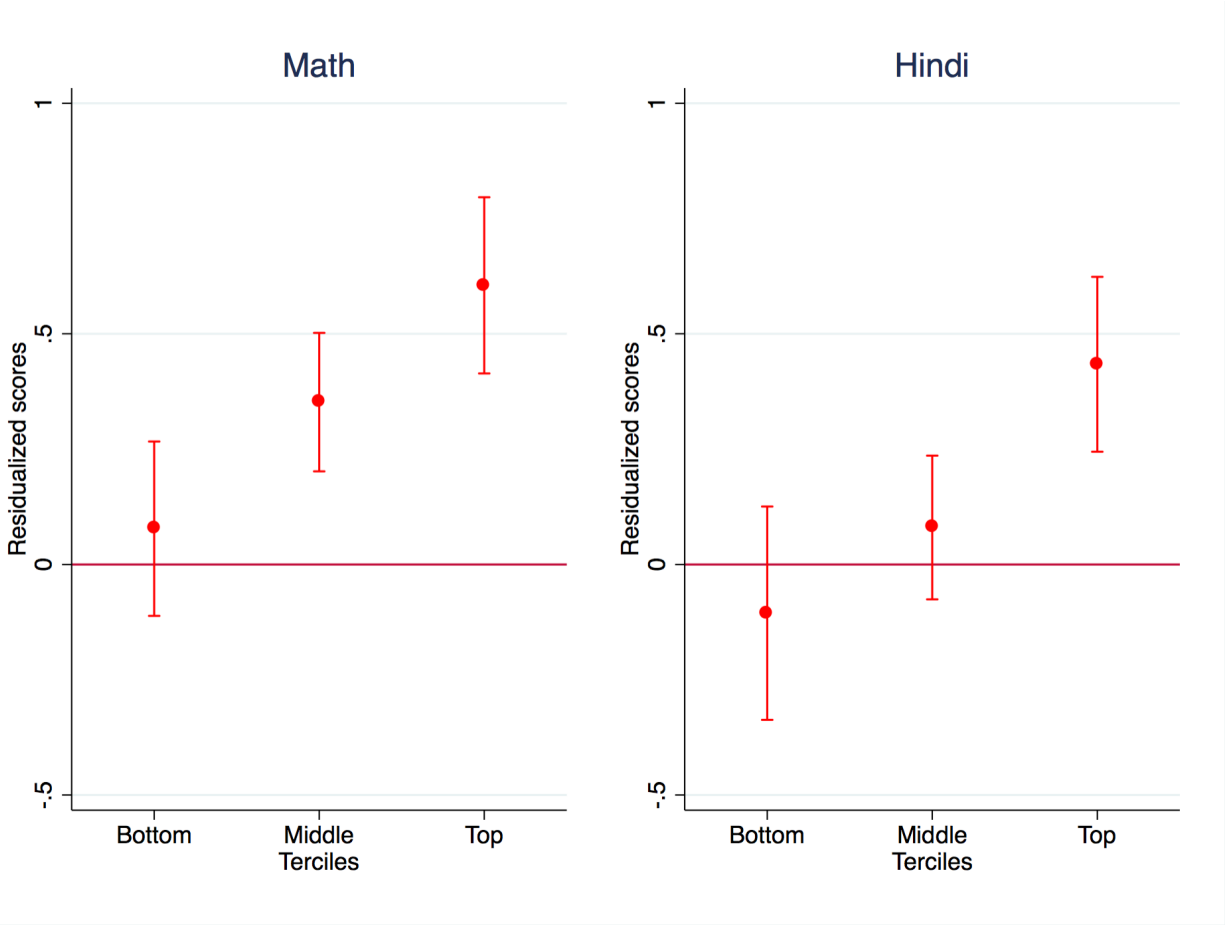
Wise, B. W. and R. K. Olson (1995). Computer-based phonological awareness and reading instruction. *Annals of Dyslexia 45*, 99–122.

Figure 1: Assessed ability levels vs. current grade enrolled in school



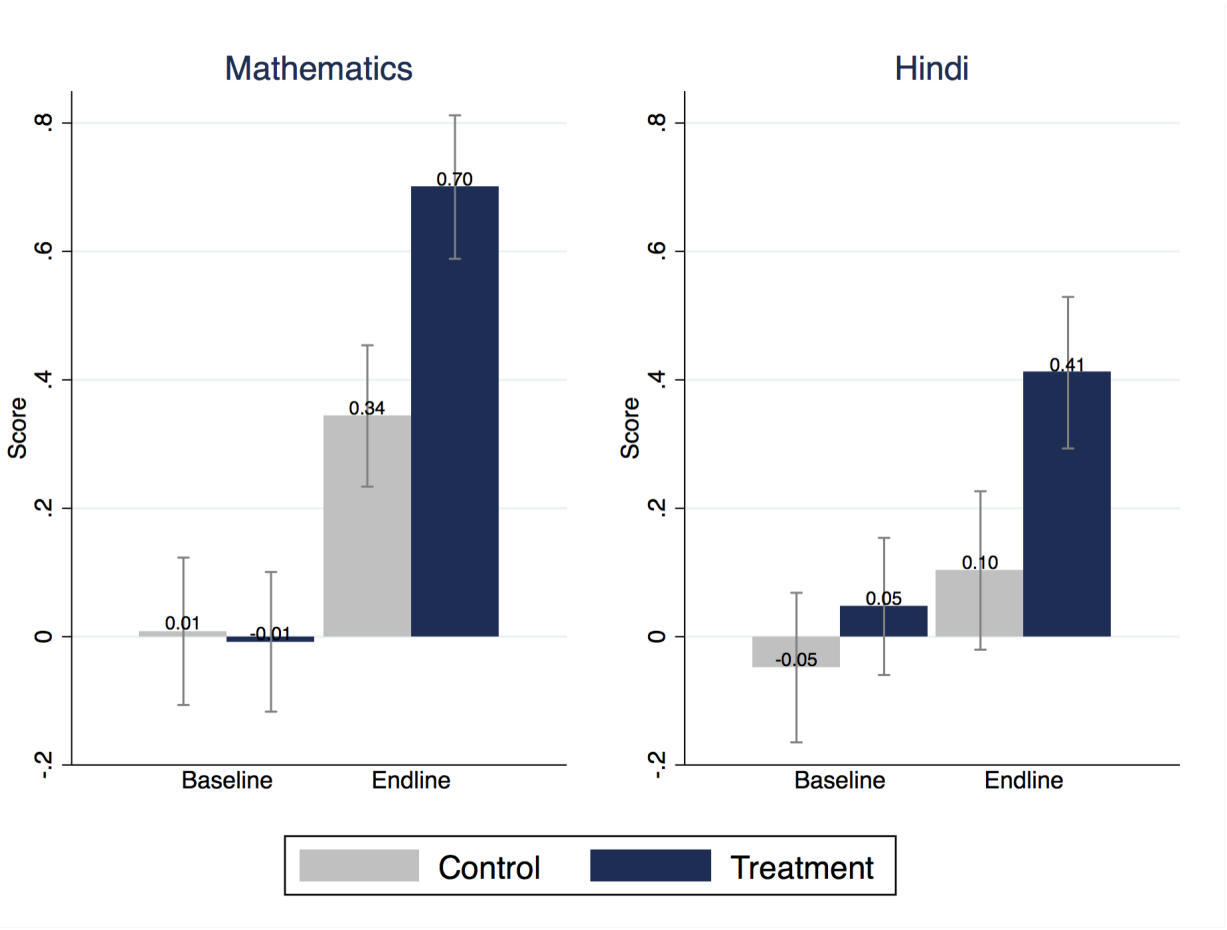
Note: This figure shows, for treatment group, the actual ability level (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. In both subjects, it shows three main patterns: (a) there is a general deficit between average attainment and grade-expected norms; (b) this 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



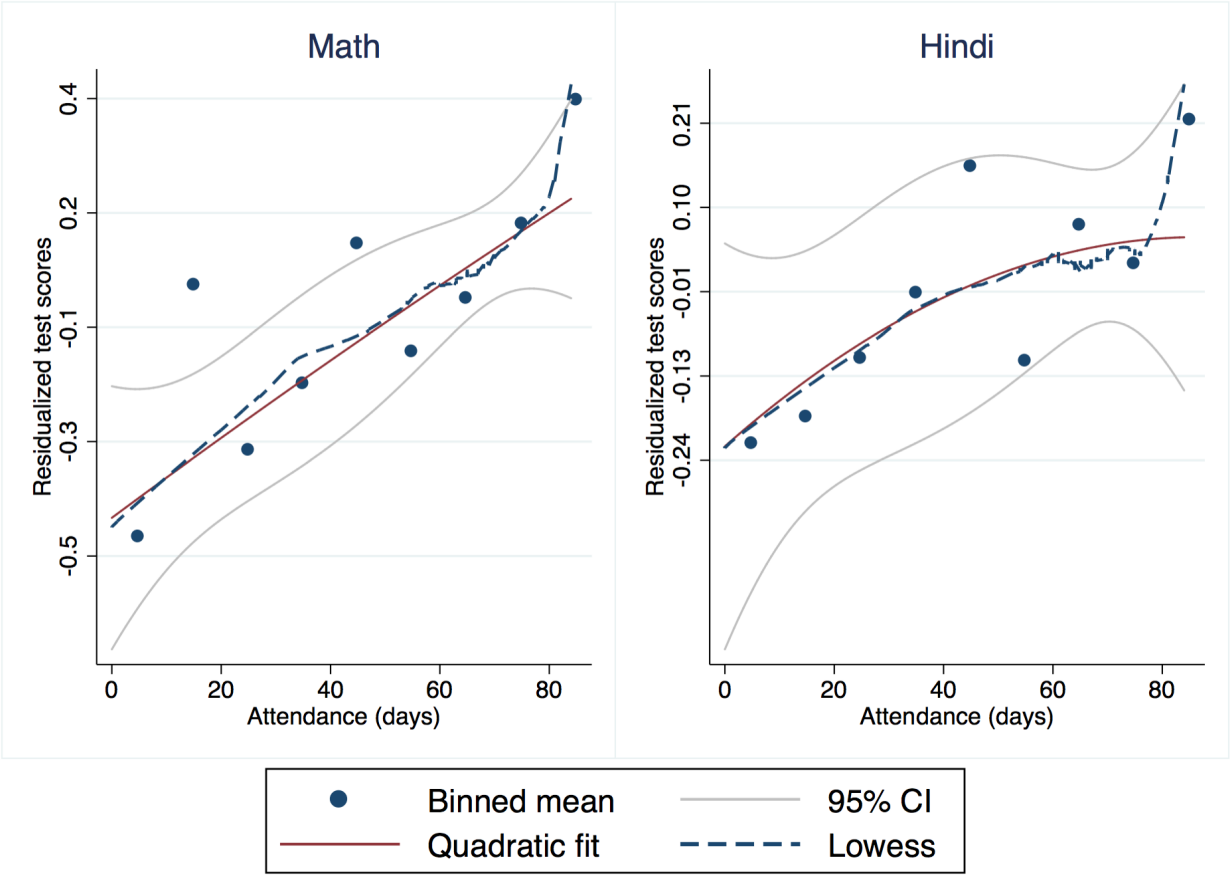
Note: This figure shows the value-added in the control group for students in different terciles of the within-grade achievement distribution. Value-added is measured on our independently-administered tests at baseline and endline tests in September 2015 and February 2016 respectively.

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



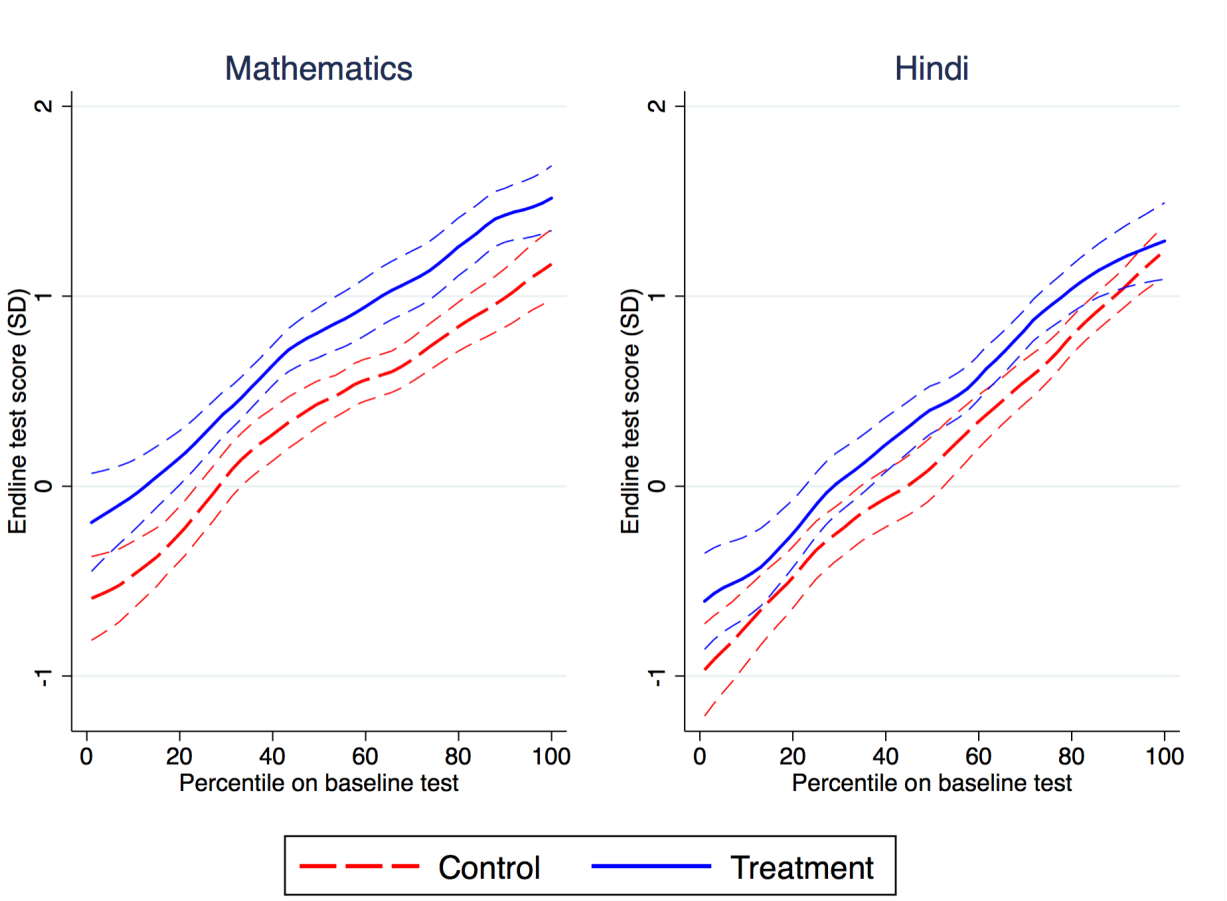
Note: This figure shows mean of test scores, normalized with reference to baseline, across treatment and control groups in the two rounds of testing with 95% confidence intervals. Test scores were linked within-subject through IRT models, pooling across grades and across baseline and endline, and are 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



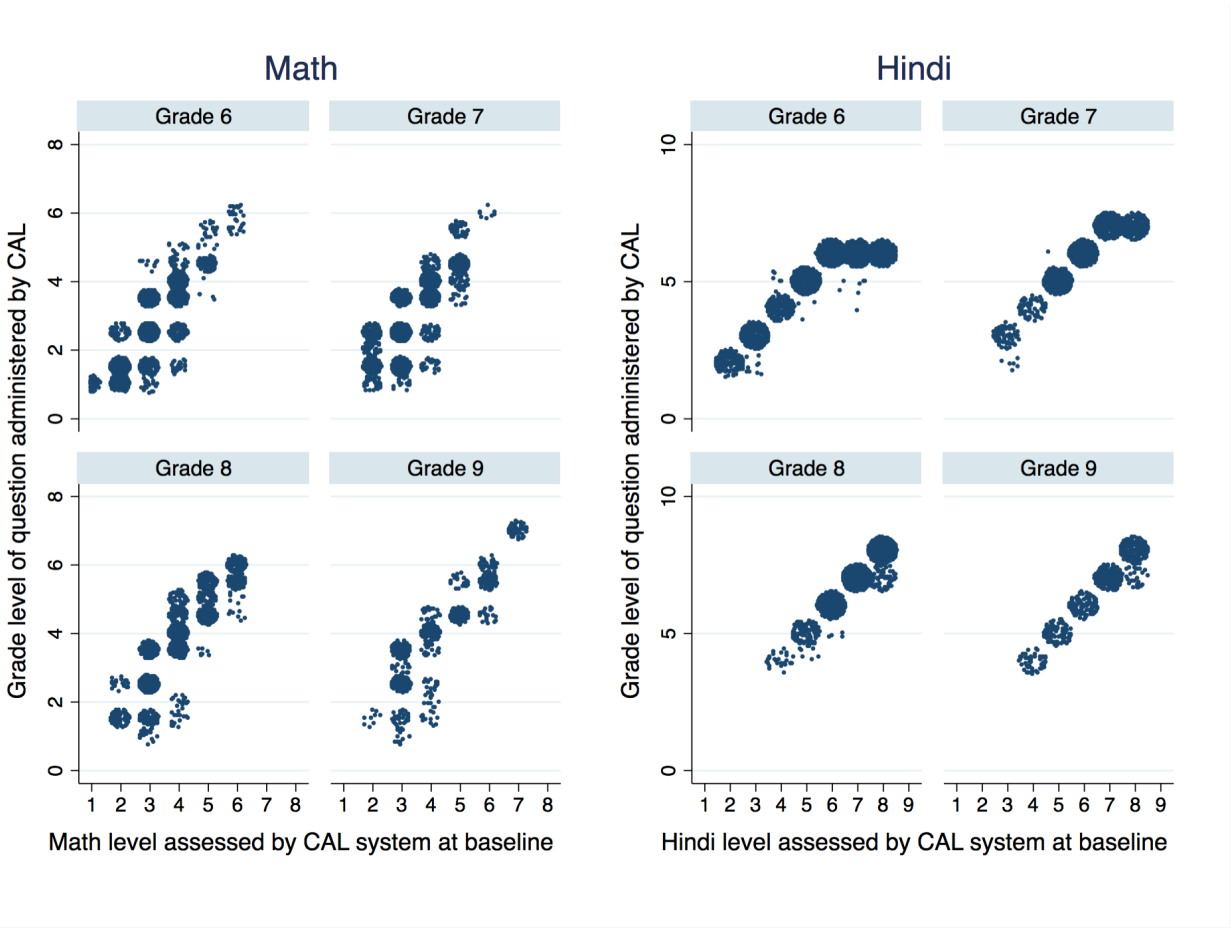
Note: This figure explores the relationship between value-added and attendance in the Mindspark program among the lottery-winners. It presents the mean value-added in bins of attendance along with a quadratic fit and a lowess smoothed non-parametric plot.

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



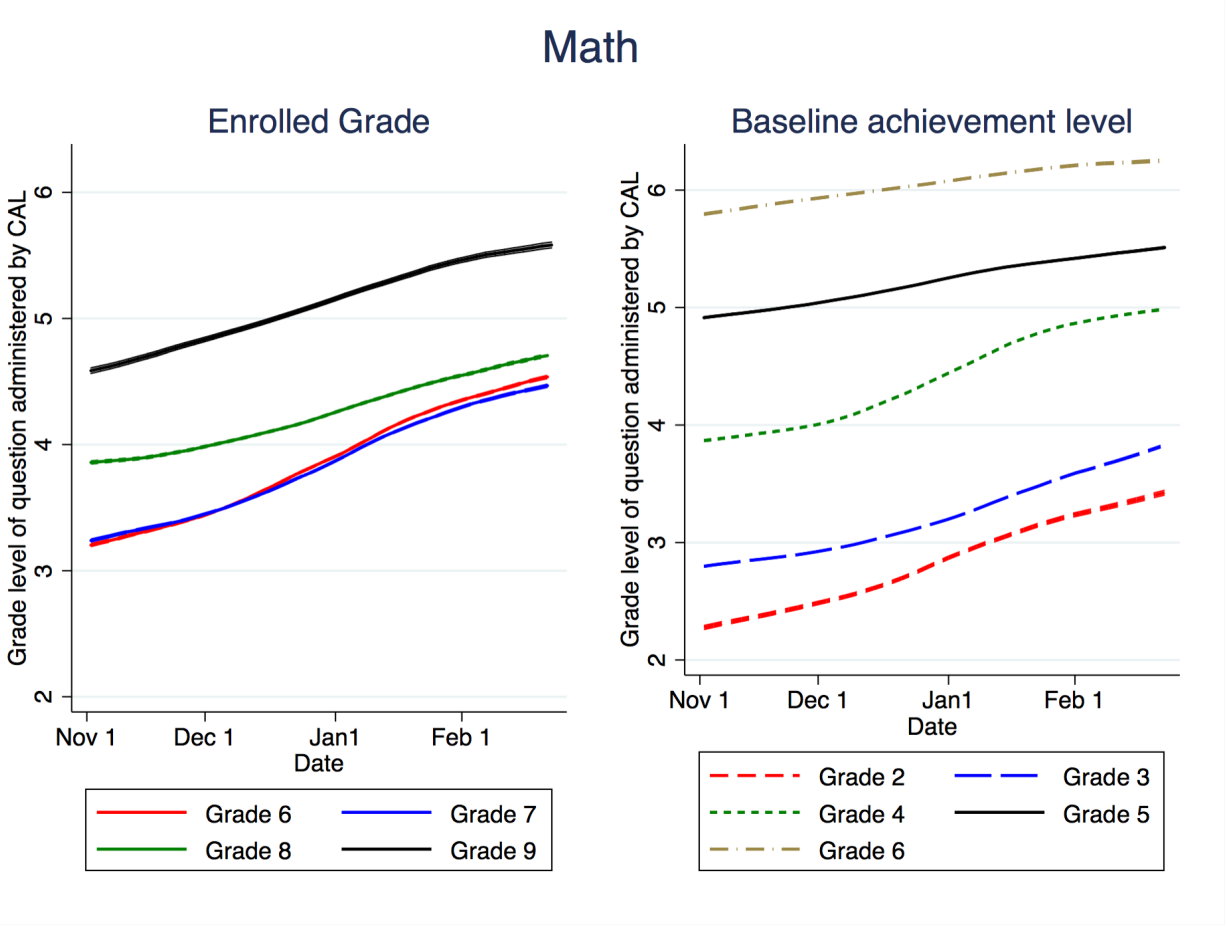
Note: The figures present local polynomial plots of degree zero (local mean smoothing) which relate endline test scores to percentiles in the baseline achievement, separately for the treatment and control groups, alongside 95% confidence intervals. At all percentiles of baseline achievement, treatment 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



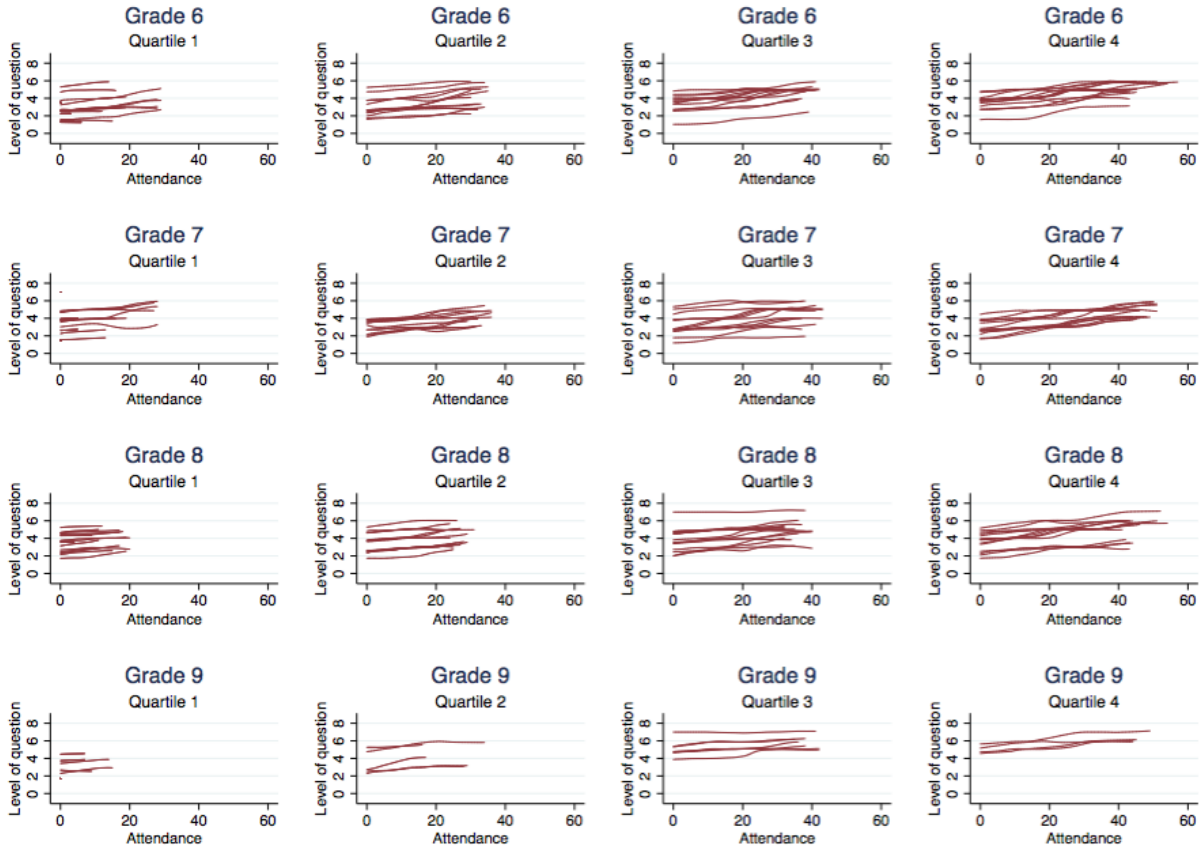
Note: This figure shows, for treatment group, the grade level of questions administered by the computer adaptive system to students on a single day near the beginning of the intervention. In each grade of enrolment, actual level of student attainment estimated by the CAL software differs widely; this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

Figure 7: Dynamic updating and individualization of content in Mindspark



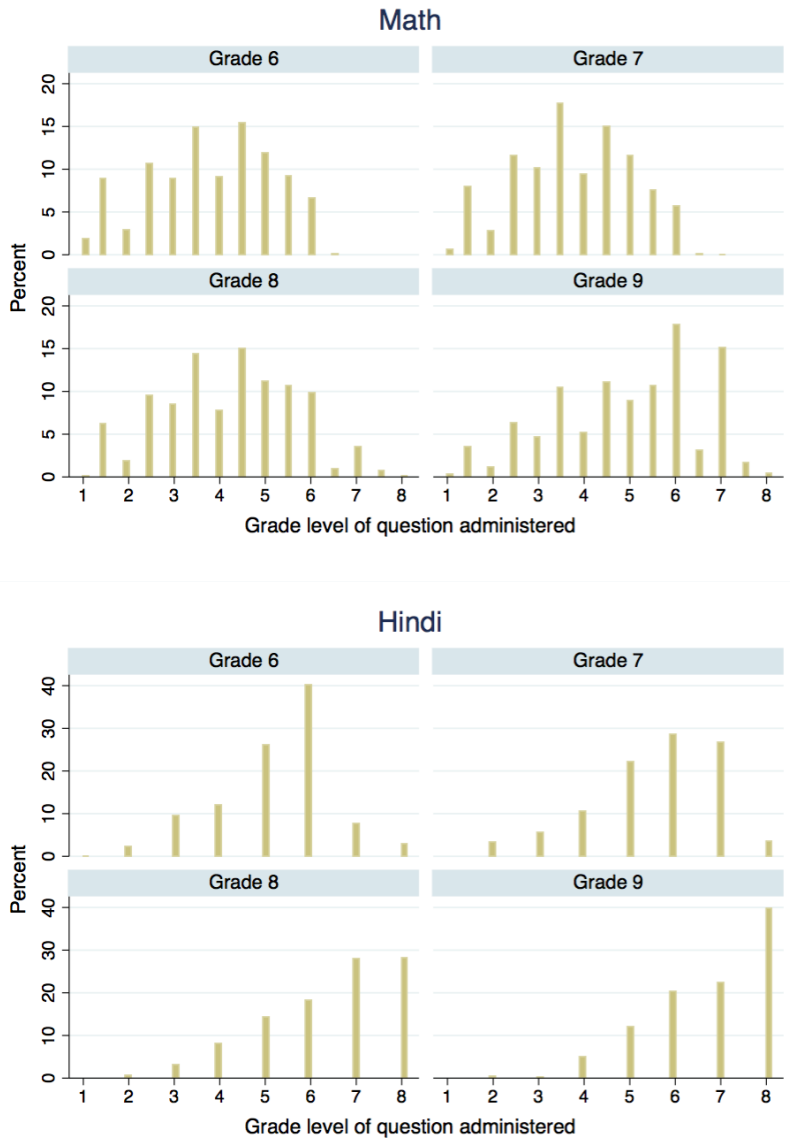
Note: This figure shows kernel-weighted local polynomial smoothed lines relating the level of difficulty of the math questions administered to students in the treatment group with the date of administration. The left panel presents separate lines by the actual grade of enrolment. The right panel presents 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: 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 9: 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 extra tuition 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 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. 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 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.

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 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. 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 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.

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. 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.

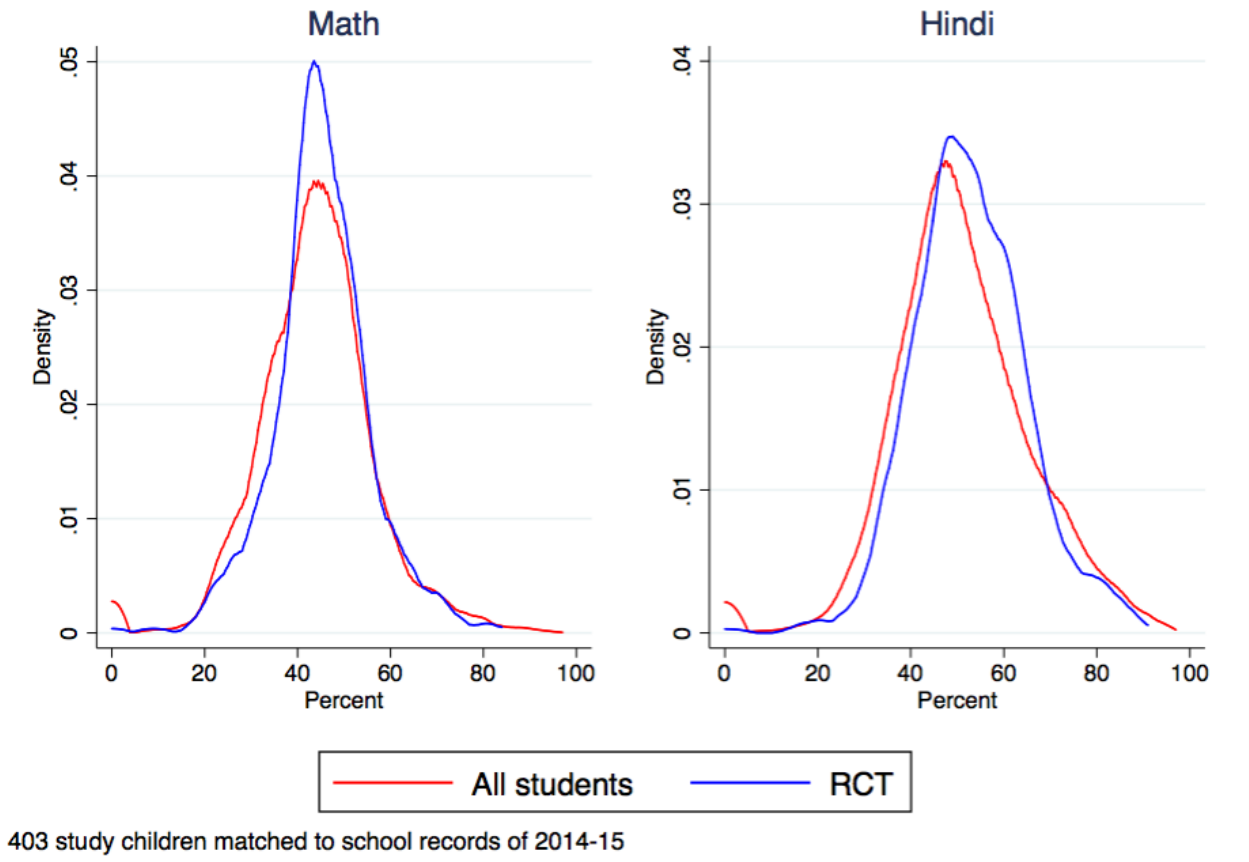
Table 9: Treatment effect on take-up of other tuition

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 tuition in various subjects. The dependent variable is whether a child was attending extra tuition 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.

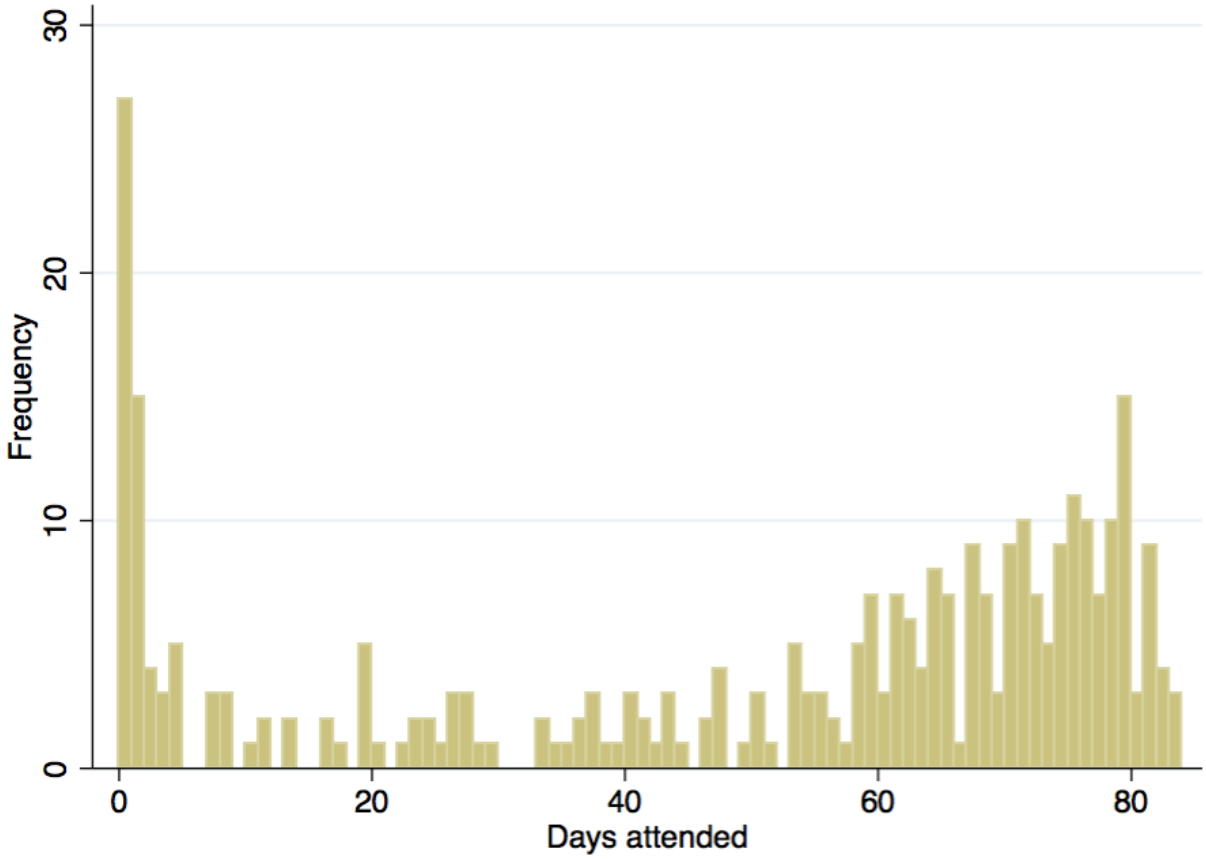
Appendix A Additional figures and tables

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.

Table A.1: Correlates of attendance

VARIABLES	(1)	(2)	(3)
	Attendance (days)		
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 math tuition			-1.95
			(4.41)
Attends Hindi tuition			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.

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.[43] Otherwise, we mentioned that a program had "no effect" on the respective subject.[44]

[43] 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.

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

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 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. (2014)	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. (2014)	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. (2015)	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	Language and reading (English)	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
Campuzano et al. (2009)	RCT	Grade 1, 12 public schools in 2 school districts, United States	Destination Reading - Course 1	Reading (English)	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 (English)	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 offers a description of how the Mindspark computer-assisted learning (CAL) software operates. This description applies only to the version of the software used at the stand-alone centers that we evaluated.

C.1 Computer training

The first time that students log into the Mindspark software, they are shown a screen that gives them the option of doing exercises on math or language. However, students can choose to skip them and proceed directly to the content sessions. The exercises take 10-15 minutes.

C.2 Diagnostic test

Once students complete the computer training, upon their first session, they are presented with a diagnostic test in the subject they selected. This test contains four to five questions per grade level on that subject. There are separate diagnostic tests for math and Hindi and the content of the test varies depending on the grade level of the student. 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.[45] 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.[46] An algorithm decides how a student's performance on the diagnostic test determines the grade level of the first set of questions he/she sees. Once a student begins interacting with the Mindspark software, the diagnostic test plays no further role.

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.

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

[46]For example, a grade 4 student who answers less than 25% of the grade 5 questions correctly will not be shown grade six questions.

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 currently has 21 topics, 105 teacher topics, 550 clusters, 11,000 SDLs, and 35,000 questions. Each teacher topic has an average of five questions, each cluster also has an average of five questions, and each SDL has an average of 20 questions. The math content is organized in this way because math learning is mostly linear (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 linear (e.g., you may still understand a text even if you do not understand grammar or all the vocabulary words in it). This is also why there are no SDLs in Hindi, and students need not pass a low-difficulty passage question before they move on to a medium-difficulty question.

C.4 Adaptability

In math, the questions within a teacher topic progressively increase in difficulty, based on EI's data analytics and 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 remedial activity which would be a step-by-step explanation of a concept or 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.

[Insert Figure C.1 here.]

Figure C.2 shows three examples of student errors in the hidden numbers game. These frequent errors were identified by the Mindspark software, and subsequently EI staff interviewed some students who made these errors to understand their underlying misconceptions. Five percent of students exhibited what EI calls "whole number thinking": they 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 separator (i.e., 3 and 27) and concluded that if 27 is greater than 3, 3.27 must be greater than 3.3. Four percent of students displayed "reverse order thinking": they believed that 3.18 was greater than 3.27 because they thought that the place value of the digits increases to the right of the decimal separator, mirroring the place value of the digits to the left of the decimal separator (e.g., hundreds, tens, ones. tens, hundreds, thousands). Therefore, they compared 81 to 27 and concluded that 3.18 must be greater than 3.27. Finally, three percent of students revealed engaging in "reciprocal thinking": they believed that 3.27 was greater than 3.39 because they mistook the numbers to the right of the decimal separator for their reciprocals, so they concluded that $3\frac{1}{27}$ was greater than $3\frac{1}{39}$. As Figure C.1 shows, the student is assigned to different activities depending on the mistake(s) he/she has made.

[Insert Figure C.2 here.]

In Hindi, in the first part, students start with passages of low difficulty and move progressively 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 move progressively towards higher-difficulty questions. (Thus, a student might be seeing low-difficulty questions on a given skill and medium-difficulty questions on another).

All decisions of the software are "hard coded" based on analyses of patterns of student errors; there is no machine-learning type algorithms.

C.5 Feedback

There is almost no direct instruction (i.e., no instructional videos). All learning happens through feedback to students on incorrect questions. Also, before each question, there is

usually an example. Additionally, 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.

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.

The software first identifies the most common errors made by students in each topic. Then, EI's education specialists use evidence from: (a) these figures; (b) interview students to enquire about the factors driving errors and misconceptions; (c) read internationally published research.

Figure C.1: Mindspark adaptability in math

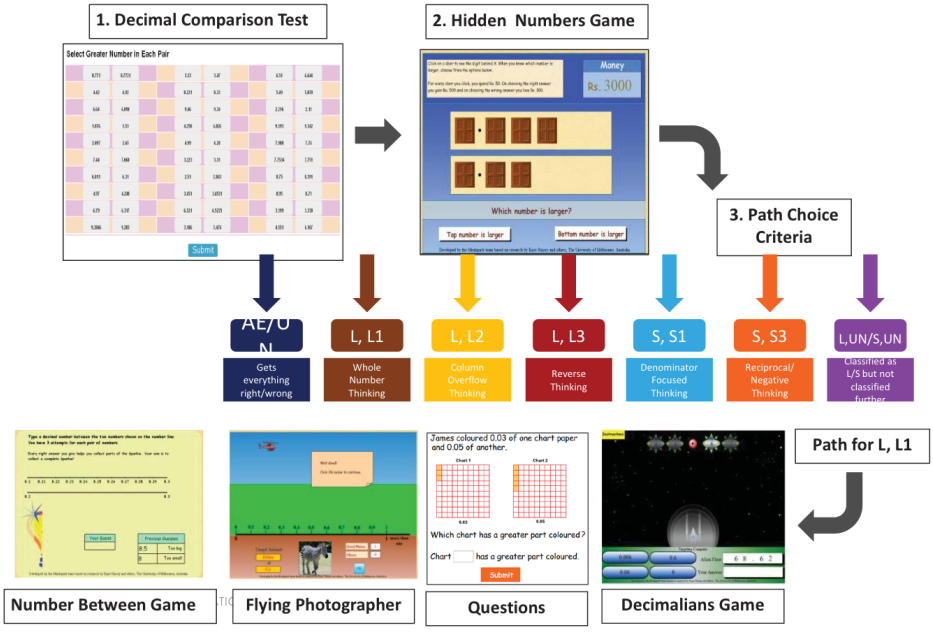
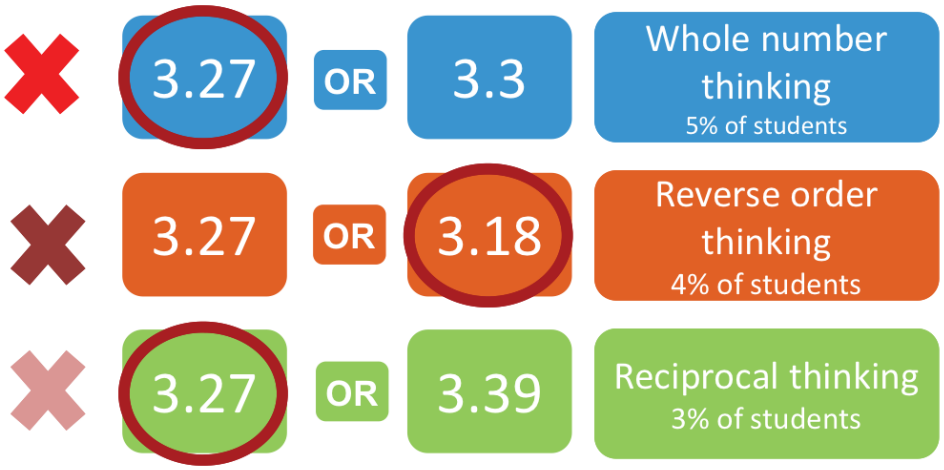


Figure C.2: Student errors in math



Appendix D Test design

D.1 Overview

Student achievement, the core outcome for this evaluation, was measured using independent assessments in math and Hindi. These were administered under the supervision of the research team at both baseline and endline. Here we present details about the test content and development, scoring and administration.

D.2 Objectives of test design

Our primary objective was to develop a test which would be informative over a wide range of ability. 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

Our focus was 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 (APREStt), 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 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. Table D.1 presents a break-up of questions by grade level of difficulty in each of the booklets at baseline and endline.

[Insert Table D.1 here.]

The assembled 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.

[Insert Table D.2 here.]

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. It is also advantageous since it then allows us to pool all test observations together in the analysis which is useful for reasons of statistical power and presentationally.

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.

In our application, reflecting the nature of the test questions, all of which were multiple-choice questions, responses to which were scored dichotomously as correct or incorrect, we used the 3-parameter logistic model to score tests. These models posit 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 which accounts for the fact that in a multiple-choice test, even an individual with no knowledge may have a non-zero probability of answering a question correctly.

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

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 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

Since a core objective of our test design and implementation protocols was to ensure that our achievement measures successfully captured the full range of student achievement in our samples, it is useful to see that raw percentage correct scores do not suffer from ceiling or floor effects. Figure A.1 presents the percentage correct responses in both math and Hindi for baseline and endline. As may be seen, the tests offer a well-distributed measure of achievement with few students unable to answer any question or to answer all questions correctly.

[Insert Figure D.1 here.]

Figure A.2 presents similar graphs for the distribution of IRT test scores. Please note that raw percentage correct test scores are not comparable over rounds or across booklets because of the different composition of test questions. IRT scores used in the analysis, distributions for which are presented in Fig A2 are comparable across rounds.

[Insert Figure D.2 here.]

D.7 Item fit

IRT models posit a parametric relationship between the underlying ability and item characteristics. An intuitive check for the performance of the IRT model is to assess the

empirical fit of the data to the estimated item characteristics. Importantly, IRT models assume that item characteristics are invariant across individuals (in the psychometrics literature, referred to as no differential item functioning).

In Figure A.3 we plot 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. As can be seen, 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.

The absence of DIF is also reassuring since it also provides suggestive evidence against "teaching to the test" in the program. Specifically, if the implementers are able to teach to the test better for some items than others (as is reasonable, given that some of the items come from tests which EI is familiar with and others from international assessments), we should have seen some evidence of DIF in these items.

[Insert Figure D.3 here.]

Figure D.1: Distribution of raw percentage correct scores

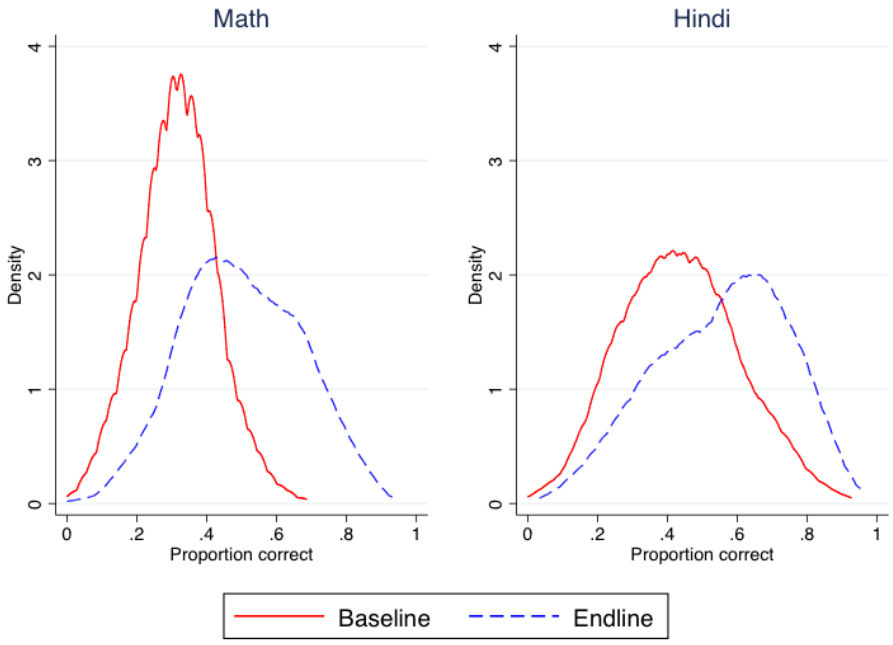


Figure D.2: Distribution of IRT scores, by round and treatment status

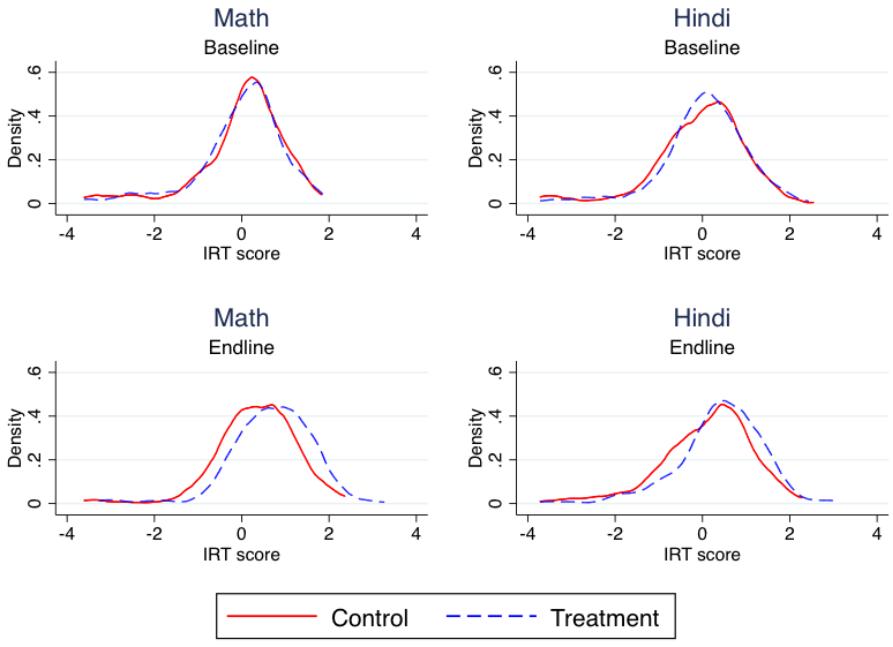
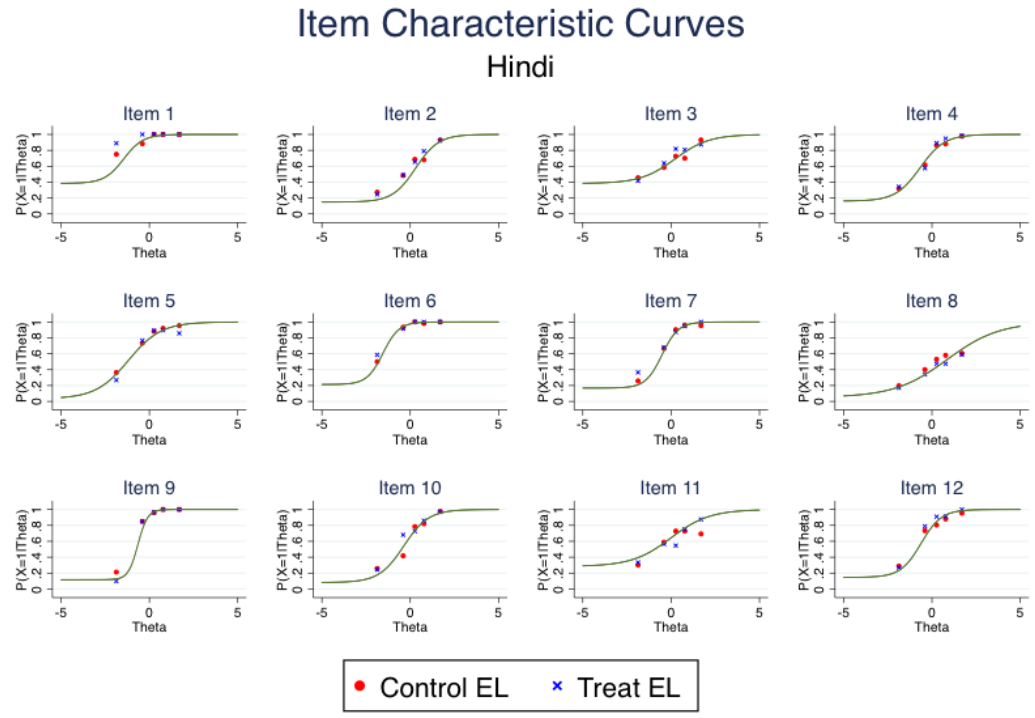
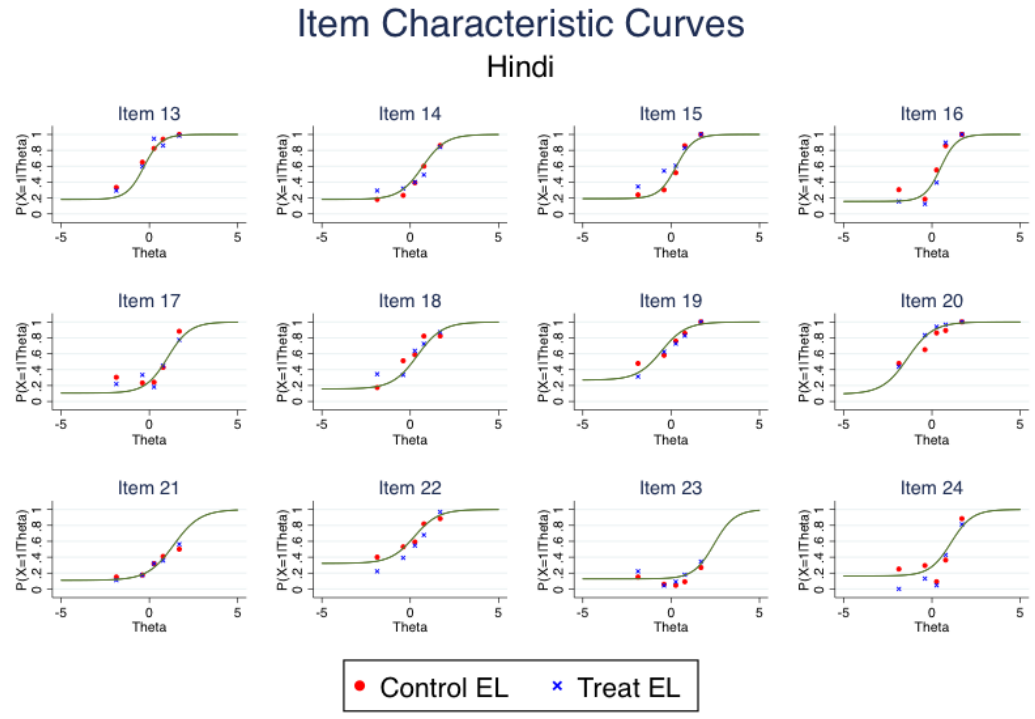


Figure D.3: Item Characteristic Curves: Hindi



Combining all grades



Combining all grades

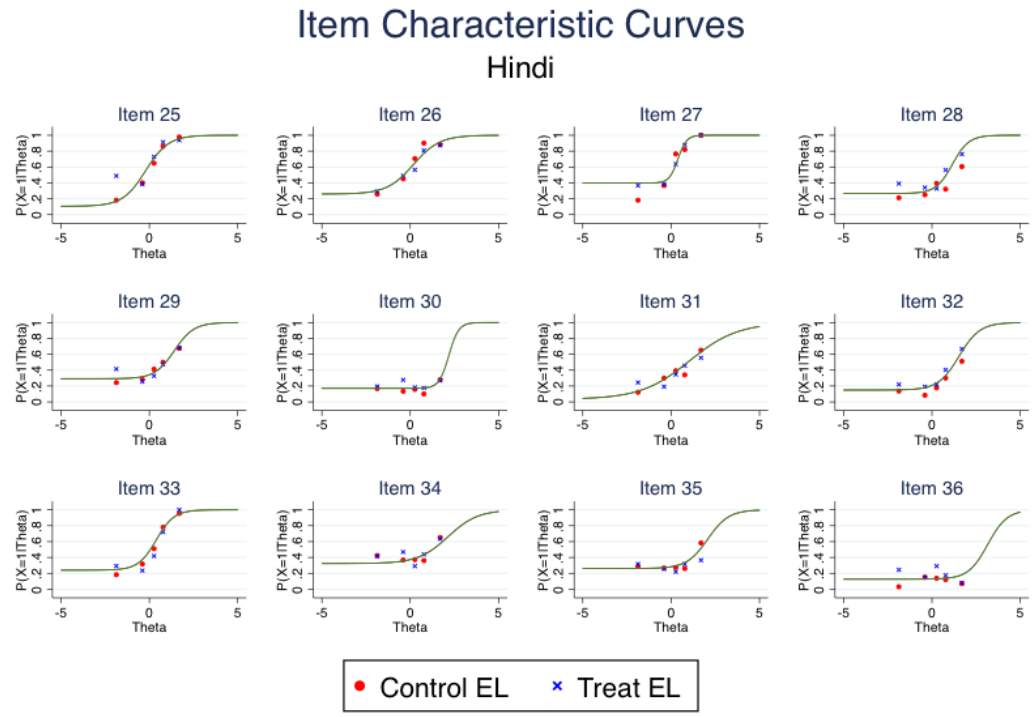
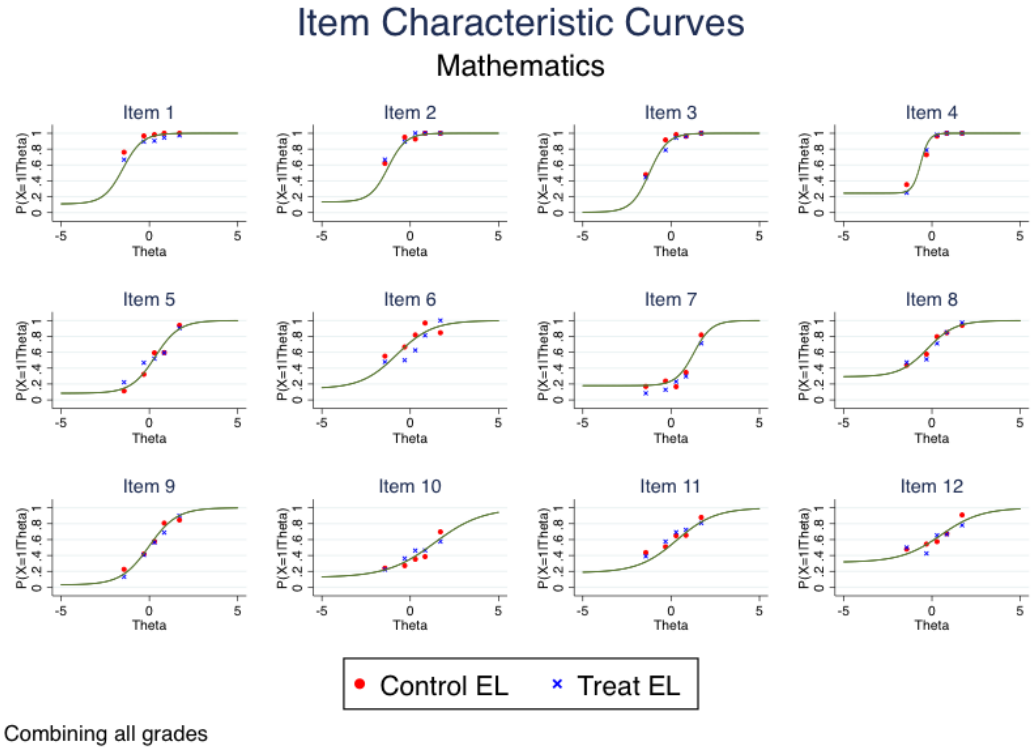
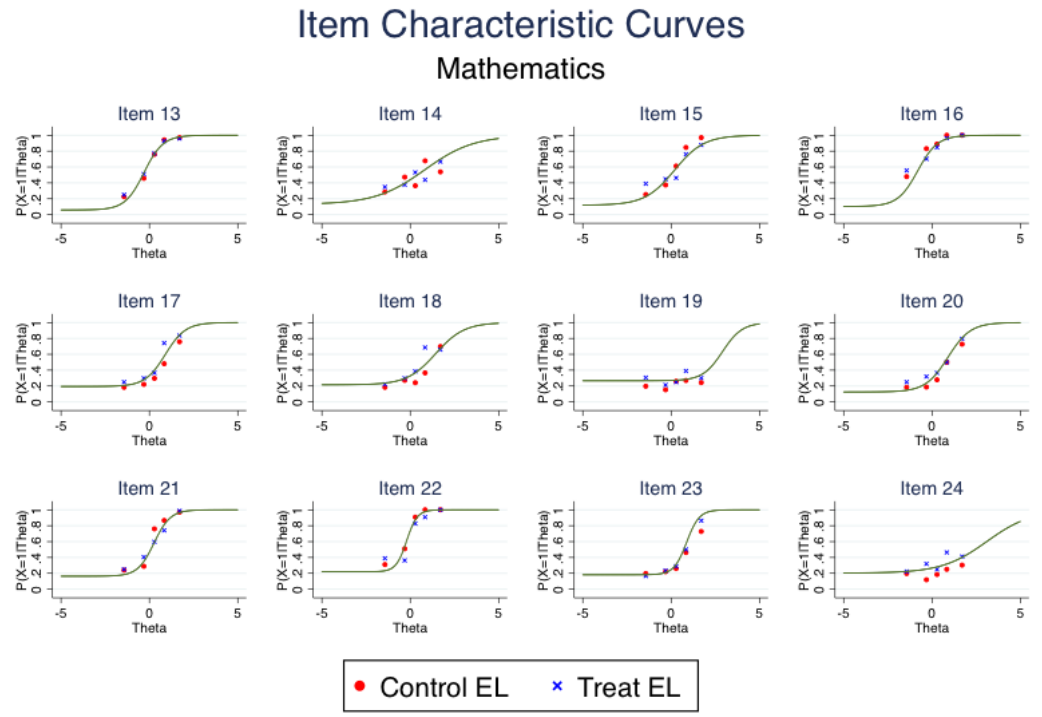
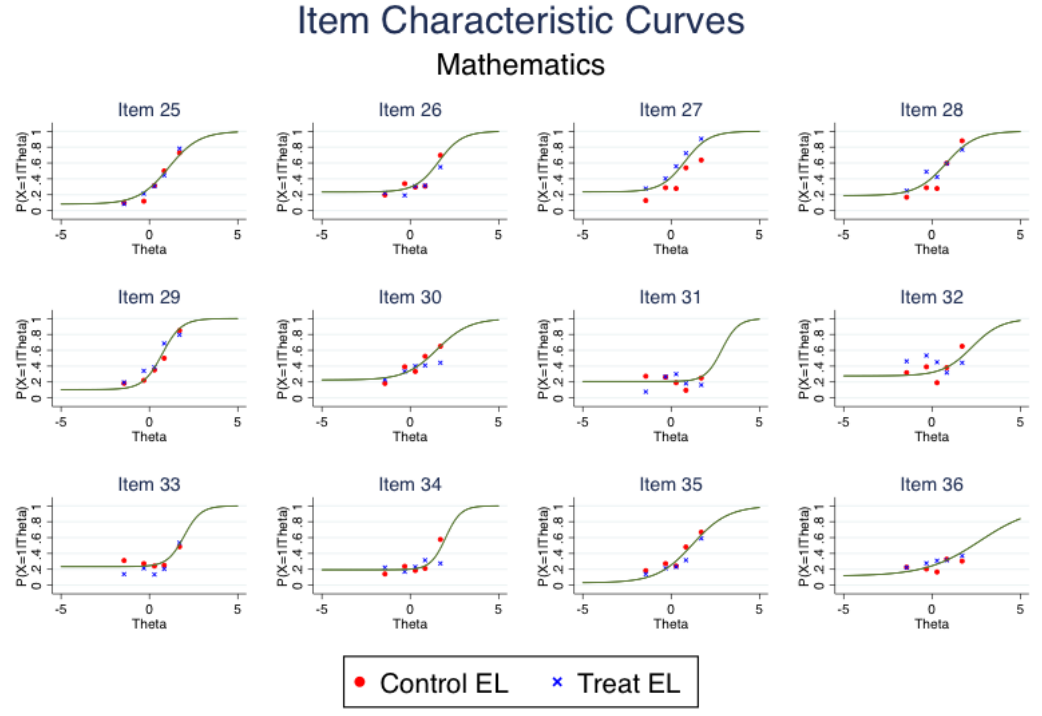


Figure D.4: Item Characteristic Curves: Math





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.