American Economic Review (Accepted September 2018)

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

Karthik Muralidharan and Abhijeet Singh and Alejandro J.Ganimian*

September 17, 2018

Abstract

We study the impact of a personalized technology-aided after-school instruction program in middle-school grades in urban India using a lottery that provided winners free access to the program. Lottery winners scored 0.37σ higher in math and 0.23σ higher in Hindi over just a 4.5-month period. IV estimates suggest that attending the program for 90 days would increase math and Hindi test scores by 0.6σ and 0.39σ respectively. We find similar absolute test score gains for all students, but much greater relative gains for academically-weaker students. Our results suggest that well-designed technology-aided instruction programs can sharply improve productivity in delivering education.

JEL codes: C93, I21, J24, O15

Keywords: computer-aided learning, productivity in education, personalized learning, teaching at the right level, post-primary education, middle school, secondary school

*Muralidharan: Department of Economics, University of California San Diego; NBER; J-PAL. E-mail:kamurali@ucsd.edu. Singh: Department of Economics, Stockholm School of Economics. E-mail: abhijeet.singh@hhs.se. Ganimian: NYU Steinhardt School of Culture, Education, and Human Development; E-mail: alejandro.ganimian@nyu.edu. We thank Esther Duflo (the editor), Abhijit Banerjee, James Berry, Peter Bergman, Prashant Bharadwaj, Gordon Dahl, Roger Gordon, Heather Hill, Priya Mukherjee, Chris Walters and several seminar participants for comments. We thank the staff at Educational Initiatives (EI)—especially, Pranav Kothari, Smita Bardhan, Anurima Chatterjee, and Prasid Sreeprakash—for their support of the evaluation. We also thank Maya Escueta, Smit Gade, Riddhima Mishra, and Rama Murthy Sripada for excellent research assistance and field support. Finally, we thank J-PAL's Post-Primary Education initiative for funding this study. The study was registered with the AEA Trial Registry (RCT ID: AEARCTR-0000980). The operation of Mindspark centers by EI was funded by the Central Square Foundation, Tech Mahindra Foundation and Porticus. All views expressed are those of the authors and not of any of the institutions with which they are affiliated.

1 Introduction

Developing countries have made impressive progress in improving school enrollment and completion in the last two decades. Yet, their productivity in converting education investments of time and money into human capital remains very low. For instance, in India, over 50% of students in Grade 5 cannot read at the second-grade level, despite primary school enrollment rates over 95% (Pratham, 2017). Similar patterns are seen in several other developing countries as well (World Bank (2018)). A leading candidate explanation for this low productivity is that existing patterns of education spending and instruction may not alleviate a key binding constraint to learning, which is the mismatch between the level of classroom instruction and student learning levels (see Glewwe and Muralidharan (2016) for a review of the evidence).

Specifically, the rapid expansion of education in developing countries has led to the enrollment of millions of first-generation learners, who lack instructional support when they fall behind the curriculum. Students who fall behind may then learn very little in school if the level of classroom instruction (based on textbooks that follow ambitious curricular standards) is considerably above their learning level (Banerjee and Duflo, 2012; Pritchett and Beatty, 2015). In Appendix B, we show that the problems of large fractions of students being behind grade-level standards, considerable heterogeneity in learning levels of students within the same grade, and mismatch between the level of student learning and the level of curriculum and pedagogy, are widespread across developing country-contexts. These problems are exacerbated at higher grades, because students are often automatically-promoted to the next grade without having acquired foundational skills. While pedagogical interventions that aim to "Teach at the Right Leve" with human support have been successful at the primary level (Banerjee et al., 2016), there is very little evidence to date on effective instructional strategies for post-primary education in developing country settings with wide heterogeneity in student learning levels.

One promising option for addressing this challenge is to make greater use of technology in instruction. While there are several mechanisms by which computer-aided learning (CAL) can improve teaching and learning,[1] a particularly attractive feature is its ability to deliver individually-customized content to "Teach at the Right Level" for all students, regardless of the extent of heterogeneity in learning levels within a classroom. However, while technology-aided instruction may have a lot of *potential* to improve post-primary education in developing countries, there is limited evidence of notable successes to date (Banerjee et al., 2013).

This paper presents experimental evidence on the impact of a technology-led instructional program (called Mindspark) that was designed to address several constraints to effective

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

pedagogy in developing countries. Reflecting over a decade of product development, a key feature of the software is that it uses its extensive item-level database of test questions and student responses to benchmark the initial learning level of every student and dynamically personalize the material being delivered to match the level and rate of progress made by each individual student. Mindspark can be delivered in a variety of settings (in schools, in after-school centers, or through self-guided study); it is platform-agnostic (can be deployed through computers, tablets, or smartphones); and it can be used both online and offline.

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

We start by presenting three key facts about the context. First, we show that the average student achievement in our sample (measured at baseline) is several grade-levels behind grade-appropriate standards and that this gap grows by grade. The average grade 6 student is around 2.5 grade levels below grade 6 standards in math; by grade 9, this deficit increases to 4.5 grade levels. Second, we show that there is considerable heterogeneity in within-grade student learning levels. Students enrolled in the same grade typically *span five to six grade levels* in their preparation, with the vast majority of them being below grade-level standards. Thus, the default of classroom instruction based on grade-appropriate textbooks is likely to be considerably above the preparation level of academically-weaker students. Consistent with this, we find that the absolute value-added on our independently-administered tests is *close to zero* for the bottom-third of students in the control group, and we cannot reject that these students made no academic progress through the school year, despite being enrolled in school.

We report four main sets of results based on the experiment. First, we find that students winning a program voucher scored 0.37σ higher in math and 0.23σ higher in Hindi relative to students who applied for but did not win the lottery. Relative to the control group, lottery winners experienced over twice the test score value-added in math and around 2.4 times that in Hindi during the study period of 4.5 months. These are intent-to-treat (ITT) estimates

[2]The teaching assistant focused on helping students with completing homework and exam preparation, while the instruction was mostly provided by the Mindspark software (see sections 2.1.1 and 5.1 for details).

reflecting an average attendance rate of 58%. Using the lottery as an instrumental variable for attendance (and additional assumptions discussed in Section 4.4), we estimate that attending the Mindspark centers for 90 days (which corresponds to 80% attendance for half a school year), would raise math and Hindi test scores by 0.6σ and 0.39σ respectively.

Second, the ITT effects do not vary by students' baseline test scores, gender, or household socioeconomic status. Thus, consistent with the promise of computer-aided learning to customize instruction for each student, the intervention was equally effective at improving test scores for *all* students. However, while the absolute impact 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 within-grade baseline test-score distribution.

Third, we examine heterogeneity of ITT effects by test-question difficulty. Since student learning levels were far below grade level in Math, the Mindspark system (which customized content to each student's learning level) mainly provided students with content at below grade-level difficulty. In Hindi, where learning gaps relative to curricular standards were smaller, students were provided with content both at and below grade-level difficulty. The test-score results reflect this pattern of instruction: In math, the test-score gains are only seen in questions of below grade-level difficulty; whereas, in Hindi test-score gains are found in questions both at and below grade-level.

Finally, we also test for ITT effects on the annual school exams. These were conducted at the school (independent of the research team) and targeted at a grade-appropriate level. Consistent with the pattern of Mindspark instruction described above, we find significant improvements in average test scores on school exams in Hindi but not in Math. We also find meaningful heterogeneity by students' initial learning level. Treated students in the lowest tercile of the within-grade baseline test-score distribution show no improvement on school tests in any subject (consistent with these students not getting exposure to any grade-level content on Mindspark). In contrast, students in the top tercile (who were more likely to receive grade-level content on the Mindspark platform) score higher in all subjects on grade-appropriate school tests as well.[3]

The test score value-added in the treatment group was over 100% greater than that in the control group, and was achieved at a lower cost per student than in the public schooling system. Thus, the program was cost effective even at the very small scale evaluated in this study, and is likely to be highly cost effective at a larger scale (since marginal costs are much lower than the average cost in our study). Further, given large learning deficits in developing countries and

[3]These results also highlight the importance of ensuring that tests used for education research are informative over a wide range of student achievement (especially in developing country settings with wide variation in within-grade student learning). Using only grade-appropriate tests (or school tests) would have led to incorrect inference regarding average program impact (see discussion in Section 4.3.3).

finite years of schooling, it is also worth considering productivity *per unit of time*. For instance, Muralidharan (2012) finds that providing individual-level performance bonuses to teachers in India led to test score gains of 0.54σ and 0.35σ in math and language after five years of program exposure. This is one of the largest effect sizes seen to date in an experimental study on education in developing countries. Yet, we estimate that regularly attending Mindspark centres could yield similar gains in one tenth the time (half a year).

The effects presented above represent a combination of the Mindspark computer-aided learning (CAL) program, group-based instruction, and extra instructional time (since we study an after-school program), and our study design does not allow us to experimentally distinguish between these channels of impact. However, a contemporaneous experimental study on the impact of an after-school group tutoring program that was also in Delhi, also targeted middle-school students, and featured an even longer duration of after school instruction found *no impact* on test scores (Berry and Mukherjee, 2016). These results suggest that extra after-school instructional time or group-based tutoring on their own may have had limited impact on student learning without the CAL program. Thus, while our experimental estimates reflect the composite impact of a 'blended learning' program, they are most likely attributable to the CAL component and not the group instruction (see discussion in section 5.1).

Our results are directly relevant to policy debates on effective strategies to address the challenge of mismatch between student learning-levels and the level of curriculum/pedagogy (which is a widespread problem in developing countries as documented in Appendix B). Many of the pedagogical interventions that have been shown to be effective in the past two decades in both South Asia and Africa have successfully addressed the challenge of mismatch by "Teaching at the Right Level" (TaRL). Practical implementation models have included providing a teaching assistant to pull out lagging students from class and teaching them basic competencies (Banerjee et al., 2007), tracking classrooms to facilitate teaching closer to the learning level of students (Duflo, Dupas and Kremer, 2011), and offering learning camps outside school hours to facilitate teaching at the right level, unencumbered by the need to complete the curriculum (Banerjee et al., 2016).

However, implementing this idea at scale is challenging for two reasons. First, most TaRL models involve either placing additional teachers in school or retraining existing teachers to conduct more differentiated instruction. This is both labor intensive, and requires considerable behavior change by existing teachers, which current evidence suggests is not easy to achieve Banerjee et al. (2016). Second, these models may not be viable at post-primary grades because the content gets more sophisticated and the extent of variation in student learning levels also increases. Our results suggest that using CAL programs like Mindspark that are able to use technology to personalize instruction to each student may provide a promising option for scaling up the TaRL approach at all levels of schooling without increasing the workload on

teachers. Further, since students can be provided differentiated instruction while maintaining the age-based cohort structure, technology-enabled personalized instruction may be able to deliver the pedagogical advantages of tracking while mitigating several of its challenges (see discussion in section 5.3).

The discussion above also helps to interpret the large heterogeneity in impacts of CAL interventions to date (see, for instance, the recent review by Bulman and Fairlie (2016)). To help place our results in the context of the existing evidence, we conducted an extensive review of existing studies with attention to the *details* of the CAL interventions that were studied (see Appendix C). Our review suggests that some clear patterns are starting to emerge. First, hardware-focused interventions that provide computers at home or at school seem to have no positive impact on learning outcomes.[4] Second, pedagogy-focused CAL programs that allow students to review grade-appropriate content at their own pace do better, but the gains are modest and range from 0.1σ to 0.2σ.[5] Finally, the interventions that deliver the largest gains (like the one we study and the one studied in Banerjee et al. (2007)) appear to be those that use technology to also personalize instruction. Thus, our results suggest that personalization (and thereby implementing TaRL) may be an important ingredient for achieving the full potential of technology-aided instruction.

More broadly, our evidence on the ability of technology-aided instruction to help circumvent constraints to human capital accumulation in developing countries, speaks to the potential for new technologies to enable low-income countries to leapfrog constraints to development. Examples from other sectors include the use of mobile telephones to circumvent the lack of formal banking systems (Jack and Suri, 2014), the use of electronic voting machines for better enfranchisement of illiterate citizens (Fujiwara, 2015) and the use of biometric authentication to circumvent literacy constraints to financial inclusion (Muralidharan, Niehaus and Sukhtankar, 2016). However, given limitations in both the ability and willingness of the poor to pay for CAL programs (see discussion in Section 5.3), government-led initiatives will likely have to play an important role in delivering on this promise.

The rest of this paper is organized as follows. Section 2 describes the intervention, and experimental design. Section 3 describes our data. Section 4 presents our main results. Section 5 discusses mechanisms, costs, and policy implications. Section 6 concludes.

[4]See, for example, Angrist and Lavy (2002); Barrera-Osorio and Linden (2009); Malamud and Pop-Eleches (2011); Cristia et al. (2012); Beuermann et al. (2015). These disappointing results are likely explained by the fact that hardware-focused interventions have done little to change instruction, and at times have crowded out student time for independent study.

[5]See, for example, Carrillo, Onofa and Ponce (2010); Lai et al. (2015, 2013, 2012); Linden (2008); Mo et al. (2014*a*); Barrow, Markman and Rouse (2009); Rouse and Krueger (2004). Anecdotal evidence suggests that pedagogy-focused CAL interventions have typically focused on grade-appropriate content in response to schools' and teachers' preference for CAL software to map into the topics being covered in class and reinforce them.

2 Intervention and Study Design

2.1 The Mindspark CAL software

Developed by a leading Indian education firm called Educational Initiatives (EI), the Mindspark software reflects over a decade of iterative product development and aims to leverage several posited channels through which education technology may improve pedagogy. At the time of the study, it had been used by over 400,000 students, had a database of over 45,000 test questions, and administered over a million questions across its users every day. The software is interactive and includes continuous student assessment alongside instructional games, videos, and activities from which students learn through explanations and feedback. We highlight some of the key design features of the software here, and provide a more detailed description with examples for each of the points below in Appendix D.

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

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

Third, even students at similar average levels of understanding of a topic, may have different specific areas of conceptual misunderstanding. Thus, the pedagogical approach needed to alleviate a student-specific conceptual 'bottleneck' may be different across students. Mindspark aims to address this issue by using its large database of millions of student-question level observations to identify patterns of student errors and to classify the type of error and target *differentiated remedial instruction* accordingly (see Appendix D.4.2 for examples). This attention to understanding patterns in student errors builds on an extensive literature in education that emphasizes the diagnostic value of error analysis in revealing the heterogeneous needs of individual students (see Radatz 1979 for a discussion). However, while the value of

error analysis is well-known to education specialists, implementing it in practice in classroom settings is non-trivial and the use of technology sharply reduces the cost of doing so.[6]

Finally, the interactive user interface, combined with the individualization of material for each student, facilitates children's *continuous engagement* with the material. The software makes limited use of instructional videos (where student attention may waver), choosing instead to require students to constantly interact with the system. This approach aims to boost student attention and engagement, to provide feedback at the level of each intermediate step in solving a problem, and to shorten the feedback loop between students attempting a problem and learning about their errors and how to correct them.

As the discussion above makes clear, Mindspark aims to use technology to simultaneously alleviate multiple constraints to effective teaching and learning in a scalable way. In future work, we hope to run micro-experiments on the Mindspark platform to try to isolate the impact of specific components of the software on learning outcomes (such as personalization, differentiated feedback, or the impact of specific pedagogical strategies). However, from the perspective of economists, we are more interested in studying the extent to which technology-aided instruction can *improve productivity* in delivering education. Thus, our focus in this paper is on studying the "full potential" impact of technology-aided instruction on education outcomes (which includes all the channels above), and we defer an analysis of the relative importance of specific components of Mindspark to future work.

2.1.1 The Mindspark centers intervention

The Mindspark CAL software has been deployed in various settings: private and government schools, after-school instructional centers and individual subscription-based use at home. Here, we evaluate the supplementary instruction model, delivered in stand-alone Mindspark centers that target students from low-income households. Students signed up for the program by selecting a 90-minute batch, outside of school hours, which they are scheduled to attend six days per week. The centers charged a (subsidized) fee of INR 200 (USD 3) per month.[7]

[6]The emphasis on error analysis reflects EI's long experience in conducting similar analyses and providing diagnostic feedback to teachers based on paper-and-pen tests (Muralidharan and Sundararaman, 2010). Thus, the Mindspark development process reflects the aim of EI to use technology to improve productivity in implementing ideas that are believed by education specialists to improve the effectiveness of pedagogy.

[7]The typical Mindspark subscription fees (in the school-based and online models) were not affordable for low-income families. Hence, the Mindspark centers were set up with philanthropic funding to make the product 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 intensity of the program, as well as the fee charged, was designed to be comparable to after-school private tutoring, typically conducted in groups of students, which is common in India. According to the 2012 India Human Development Survey, 43% of 11-17 year olds attended paid private tutoring outside of school.

Scheduled daily instruction in Mindspark centers was divided into 45 minutes of computer-based instruction and 45 minutes of supervised instructor-led group-based study. In the time allotted to the computer-based instruction, each student was assigned to a Mindspark-equipped computer with headphones that provided him/her with activities on math, Hindi and English. Two days of the week were designated for math, two days for Hindi, one day for English, and students could choose the subject on one day each week.

The group-based instruction component included all students in a given batch (typically around 15 students) and was supervised by a single instructor. Instructors were locally hired and were responsible for monitoring students when they are working on the CAL software, providing the group-based instruction, facilitating the daily operation of the centers, and encouraging attendance and retention of enrolled students.[8] Instruction in the group-based component consisted of supervised homework support and review of core concepts of broad relevance for all children without individual customization.

Thus, the intervention provided a 'blended learning' experience that included personalized one-on-one computer-aided instruction along with additional group academic support by an instructor. As a result, all our estimates of program impact and cost effectiveness are based on this composite program. Further, to the extent that the presence of an adult may be essential to ensure student adherence to the technology (both attendance and time on task), it may not be very meaningful to try to isolate the impact of the technology alone. In section 5.1, we discuss results from a parallel experimental evaluation in the same context showing *no impact* on student learning from an after-school group tutoring program (with no technology). Hence, one way to interpret our results is as an estimate of the extent to which using technology *increased the productivity of an instructor*, as opposed to technology by itself.

2.2 Sample

The intervention was administered in three Mindspark centers in Delhi focused on serving low-income neighborhoods. The sample for the study was recruited in September 2015 from five public middle schools close to the centers. All five schools had grades 6-8, three of these schools had grade 9, and only two had grades 4-5. Three were all-girls schools and the other two were all-boys schools. Therefore, our study sample has a larger share of girls in grades 6-8. In each school, staff from EI and from J-PAL South Asia visited classrooms from grades 4-9 to introduce students to the Mindspark centers and to invite them and their parents to a demonstration at the nearby center (information flyers were provided to share with parents).

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

At the demonstration sessions, students and their parents were introduced to the program and study by EI staff. Parents were told that, if their child wanted to participate in the study, he/she would need to complete a baseline assessment and that about half of the students would be chosen by lottery to receive a voucher which would waive the usual tuition fees of INR 200 per month until February 2016 (i.e. for nearly half of the school year). Students who were not chosen by lottery were told that they would be provided free access to the centers after February 2016, if they participated in an endline assessment in February 2016. However, lottery losers were not allowed to access the program during the study period. These two design features helped to reduce attrition, and increase statistical power respectively.

Our study sample comprises the 619 students who completed the baseline tests and surveys. About 97.5% of these students were enrolled in grades 6-9.[9] To assess the representativeness of our self-selected study sample (and implications for the external validity of our results), we compare administrative data on school final-exam scores in the preceding school year (2014-15) across study participants and the full population of students in the same schools. Study participants have modestly higher pre-program test scores (of around 0.15σ) than non-participants (Table A.1). However, there is near-complete common support in the pre-program test-score distribution of participants and non-participants (Figure A.1), suggesting that our results are likely to extend to other students in this setting (especially since we find no heterogeneity in impact by baseline test scores; see Section 4.3).

2.3 Randomization and Compliance

The 619 participants were individually randomized into treatment and control groups with 305 students in the control and 314 in the treatment group. Randomization was stratified by center-batch preferences.[10] The treatment and control groups did not differ significantly at baseline on gender, SES, or baseline test scores (Table 1, Panel A).[11] Of the 314 students offered a voucher for the program, the mean attendance rate was 58% (around 50 days out of a maximum possible of 86 days). The full distribution of attendance among lottery-winners is presented in Figure A.2, and we present both ITT estimates of winning the lottery and IV estimates of the dose-response relationship as a function of days of attendance in Section 4.

Of the 619 students who participated in the baseline test, 539 (87%) also attended the endline test. The follow-up rate was 85% in the treatment group and 90% in the control group. This

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

[10] Students were asked to provide their preferred slots for attending Mindspark centers given school timings and other commitments. Since demand for some slots was 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] The difference in age is significant at the 10% level (p=0.07), but this is one of several comparisons. The age variable also has more missing data since these were filled out in self-reported surveys.

difference is significant at the 10% level and so we will present inverse probability weighted estimates of treatment effects as well as Lee (2009) bounds of the treatment effect (section 4.5.1). We also find no significant difference between treatment and control groups in mean student characteristics (age, gender, SES, or baseline test scores) of those who attend both the baseline and endline test, and comprise our main study sample (Table 1, Panel B).

3 Data

3.1 Student achievement

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

The tests were designed independently by the research team and intended to capture a wide range of student achievement. Test items ranged in difficulty from "very easy" questions designed to capture primary school level competencies much below grade level to "grade-appropriate" competencies found in international assessments. Test scores were generated using Item Response Theory models to place all students on a common scale across the different grades and across baseline and endline assessments. The common scale over time allows us to characterize the *absolute* test score gains made by the control group between the two rounds of testing. The assessments performed well in capturing a wide range of achievement with very few students subject to ceiling or floor effects. Details of the test design, scoring, and psychometric properties of individual test questions are provided in Appendix E.

3.2 Mindspark CAL system data

The Mindspark CAL system logs all interactions that each student has with the software platform. This includes attendance, content presented, answers to each question presented, and the estimated grade level of student achievement at each point in time. These data are available (only) for the treatment group. We use these data in three ways: to describe the mean and distribution of learning gaps relative to curricular standards in each grade at baseline; to demonstrate the personalization of instruction by Mindspark; and to characterize the evolution of student achievement in the treatment group over the period of the treatment.

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

3.3 School records

At the school level, we collected administrative records on test scores on school exams of all students in the experiment and their peers in the same schools and classrooms. This was collected for both the 2014-15 school year (to compare the self-selected study sample with the full population of students in the same schools) and the 2015-16 school year (to evaluate whether the treatment affected test scores on school exams).

3.4 Student data

At the time of the baseline assessment, students answered a self-administered written student survey which collected basic information about their socio-economic status, and household characteristics. A shorter survey of time-varying characteristics was administered at endline. We also conducted a brief telephone survey of parents in July 2016 to collect data on use of private tutoring, and their opinion of the Mindspark program.

4 Results

4.1 Learning levels and variation under the status-quo

Data from the Mindspark CAL system provides an assessment of the actual grade level of each student's learning level regardless of grade enrolled in. We use these data to characterize learning levels, gaps, and heterogeneity among the students in our sample. The main results are presented in Figure 1, which shows the full joint distribution of the grades students were enrolled in and their assessed learning level at the start of treatment.[13]

We highlight three main patterns in Figure 1. First, most children are already much below grade level competence at the beginning of post-primary education. In grade 6, the average student is about 2.5 grades behind in math and about half a grade behind in Hindi.[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 considerably flatter than the line of equality between curricular standards and actual achievement levels. This suggests that average student academic achievement is progressing at a lower rate than envisaged by the curriculum — by grade 9, students are (on average) nearly 4.5 grades behind in math and 2.5 grades behind in Hindi. Third, the figure presents a stark illustration of the very wide

[13]Note that these data are only available for students in the treatment group. However, Figure 1 uses data from the *initial* diagnostic test, and does not reflect any instruction provided by Mindspark.

[14]While most patterns across grades are similar in the two subjects, the computer system's assessment on grade-level competence of children may be more reliable for math than for language (where competencies are less well-delineated across grades). Baseline test scores on our independent tests in both subjects are higher for students assessed by the CAL program as being at a higher grade level of achievement, which helps to validate the grade-level benchmarking by the CAL program (See Figure A.3). Further details of the diagnostic test and benchmarking by the software are presented in Appendix D.

dispersion in achievement among students enrolled in the *same* grade: students in our sample span 5-6 grade levels in each grade.

In Appendix B, we present additional evidence to show that the patterns documented in Figure 1 are likely to hold in a wide variety of developing country settings. Specifically, we show using additional datasets that (a) the wide distribution of learning levels within a single grade are also seen in other settings and (b) that a substantial proportion of students in Grade 5 (towards the end of lower primary schooling in most countries) are often as much as three grade levels behind the level expected by the curriculum. In the case of India (where we have exactly comparable data from other states), we show that both dispersion in learning levels, and the lag relative to curricular norms, are even more severe in larger representative samples in the states of Madhya Pradesh and Rajasthan, than in our study sample in Delhi.

4.2 Program Effects (Intent-to-treat estimates)

The main treatment effects can be seen in Figure 2, which presents mean test scores in the baseline and endline assessments in math and Hindi for lottery-winners and losers. While test scores improve over time for both groups, endline test scores are significantly and substantially higher for the treatment group in both subjects.

We estimate intent-to-treat (ITT) effects of winning the lottery (β) using:

$$Y_{iks2} = \alpha_s + \gamma_s.Y_{iks1} + \beta_s.Treatment_i + \boldsymbol{\phi_k} + \epsilon_{iks2} \quad (1)$$

where Y_{ikst} is student i's test score, in randomization stratum k, in subject s at period t (normalized to μ=0, σ=1 on the baseline test); $Treatment$ is an indicator variable for being a lottery-winner; $\boldsymbol{\phi}$ is a vector of stratum fixed effects; and ϵ_{iks2} is the error term.[15]

We find that students who won the lottery to attend Mindspark centers scored 0.37σ higher in math and 0.23σ higher in Hindi compared to lottery losers after just 4.5 months (Table 2: Cols. 1-2). In Cols. 3 and 4, we omit strata fixed effects from the regression, noting that the constant term α in this case provides an estimate of the absolute value-added (VA) in the control group over the course of the treatment.[16] Expressing the VA in the treatment group ($\alpha + \beta$) as a multiple of the control group VA (α), our results indicate that lottery-winners made over twice the progress in math, and around 2.4 times the progress in Hindi, compared to lottery-losers.

[15]We use robust Huber-White standard errors throughout the paper rather than clustered standard errors because of the individual (as opposed to group) randomization of students to treatment status. Common shocks from test day and venue effects are netted out by the inclusion of strata fixed effects since all students in the same stratum (both treatment and control), were tested on the same day in the same location.

[16]This interpretation is possible because the baseline and endline tests are linked to a common metric using Item Response Theory. This would not be possible if scores were normalized within grade-subject-*period* as is common practice. Note that treatment effects are very similar (0.38σ in math and 0.23σ in Hindi) when test scores are normalized relative to the within-grade distribution in the control group at the endline (Table A.2).

These are ITT results based on an average attendance of about 58% among lottery-winners. We present IV results and estimates of a dose-response relationship in Section 4.4.

In addition to presenting impacts on a normalized summary statistic of student learning, we also present impacts on the fraction of questions answered correctly on different domains of subject-level competencies (Table 3). The ITT effects are positive and significant across all domains of test questions. In math, these range from a 12% increase on the easiest type of questions (arithmetic computation), determined by the proportion correctly answered in the control group, to a 38% increase on harder competencies such as geometry and measurement. Similarly, in Hindi, ITT effects range from a 6.4% gain on the easiest items (sentence completion) to a 17% gain on the hardest competence (answering questions based on interpreting and integrating ideas and information from a passage).

4.3 Heterogeneity

4.3.1 Heterogeneity by student characteristics

We investigate whether ITT effects vary by gender, socio-economic status, or initial test scores, using a linear interaction specification and find no evidence of heterogeneity on these dimensions (Table 4). Since baseline test scores are a good summary statistic of prior inputs into education, we also present non-parametric estimates of the ITT effect as a function of baseline scores. We do this by plotting kernel-weighted locally-smoothed means of the endline test scores at each percentile of the baseline test-score distribution, separately for the treatment and control groups (Figure 3). In both math and Hindi, we see that the test scores in the treatment group are higher than those in the control group at every percentile of baseline test scores, and that the gains appear similar at all percentiles.

Next, we test for equality of treatment effects at different points of the *within-grade* test-score distribution. We do this by regressing endline test scores on the baseline test scores, indicator variables for treatment and for within-grade terciles at baseline, and interaction terms between the treatment variable and two terciles (the regression is estimated without a constant). We see limited evidence of heterogeneity here as well (Table 5). The coefficient on the treatment dummy itself is statistically significant, but the interaction terms of treatment with the tercile at baseline are typically not significant.[17]

Note, however, that we see considerable heterogeneity in student progress by initial learning level *in the control group*. While students in the top third of the baseline test-score distribution show significant academic progress between baseline and endline, it is striking that we cannot reject the null of *no increase* in test scores for the bottom-third of students in the control

[17]Point estimates suggest that treatment effects in Hindi were higher for the weakest students, but only one of the two interactions (with the middle-tercile) is significant, and the coefficient on a linear interaction between treatment and within-grade tercile is not significant (not shown).

group over the same period (with coefficients close to zero in both subjects) suggesting that lower-performing students make no academic progress under the status quo (Figure 4).

Thus, winning a voucher appears to have benefited students at all parts of the achievement distribution fairly equally, suggesting that the Mindspark software could teach all students equally well. However, since students in the lowest tercile of the within-grade baseline test score distribution did not make any academic progress in the control group on either subject, the *relative* gains from the treatment (measured as a multiple of what students would have learnt in the absence of treatment) were much larger for the weaker-performing students even though absolute gains are similar across all students (Figure 4).

4.3.2 Heterogeneity by test characteristics

Personalized instruction, combined with substantial heterogeneity in student preparation (Figure 1) may result in students with different initial learning levels gaining competences of varying difficulty. We directly test for this possibility below. We start by using the CAL system data to examine the grade-level distribution of content presented by the software to students in the treatment group (see Figure A.4). In math, most of the content presented to students by Mindspark was below grade level, with very little content at the level of the grade in which the student is enrolled. However, in Hindi, in addition to lower-grade content, a substantial portion of the Mindspark instruction in each grade was at grade level.

We find heterogeneity in test-score impacts by test characteristics consistent with the pattern of instruction on the CAL platform described above. Table 6 presents separate estimates of treatment effects on the proportion of test questions answered correctly at and at below grade level.[18] We see that while there were large treatment effects in math on items below grade level, there was *no impact on grade-level questions*. In Hindi, on the other hand, we find that the treatment effect is significant for both questions at and below grade level.

These patterns in our data are also replicated in the independent data we collected on test scores on school exams. Table 7 presents the treatment effect of being offered a voucher on scores on the annual end of year school exams held in March 2016.[19] Mirroring the results on grade-level items on our own tests, we find a significant increase in test scores of 0.19σ in Hindi but no significant effect on math. We also do not find significant effects on the other subjects (science, social science, or English), although all the point estimates are positive.

[18]Items on our tests, which were designed to capture a wide range of achievement, were mapped into grade-levels with the help of a curriculum expert.

[19]In Delhi, test papers for the annual exam are common across schools for each subject in each grade. In our regressions, we normalize test scores to μ=0, σ=1 in each grade/subject in the control group.

4.3.3 Interaction between test characteristics and student preparation

While the mean impact on school tests is not significant, students with higher baseline test scores may be more likely to also improve on (grade-level) school tests because they would be more likely to receive grade-level content on the Mindspark system. We test for this possibility and find consistent evidence that test scores also improve on school exams for treated students in the top third of the baseline test-score distribution (Table 8). For these students, test scores on school exams are higher on *every subject* (with treatment effects ranging between 0.2-0.5σ), with gains on four out of five subjects being significant (Hindi, Math, English, and Social Studies). Averaged across subjects, these students scored 0.33σ higher (p=0.03). In contrast, we find no improvements in school exam scores for the bottom two-thirds of students.[20]

We test for similar patterns on our own tests (Table A.3), and the math results are consistent with those found on the school tests: treated students in the top tercile perform better on items at grade-level (p=0.08) while students in the bottom two terciles show no program effect. However, reflecting the large deficits in math knowledge in comparison to the curriculum, treated students in all terciles make progress on below-grade items (where the treatment effect is positive and statistically significant for all terciles).[21]

These results illustrate the importance of conducting education research with well-calibrated tests that are informative over a wide range of student achievement (especially in developing country settings with wide variation in within-grade student learning). In our case, relying on grade-level assessments would have led to incorrect inference regarding program impacts, and have led to a conclusion that the program had no impact on math despite the very large gains in test scores seen on a properly constructed test. See Appendix E for further details on test design for our study, and Muralidharan (2017) for a detailed discussion on test construction for education research in general.

4.4 IV estimates of dose-response relationship

All the results presented so far are ITT estimates, which are based on an average attendance of about 58% among lottery-winners.[22] In this section, we present LATE estimates of the impact of

[20]Indeed, for the bottom-third of students, the coefficient is often negative (although typically not statistically significant). This suggests that the program, by focusing on concept-level mastery pitched at the students' achievement levels, may have crowded out other activities (such as rote memorization and practising past exam questions) that could lead to higher performance on school exams in the short term.

[21]On our tests, gains in Hindi are larger (and only statistically significant) for the bottom tercile (Table A.3). This is in contrast to the school results, where the gains are larger (and only statistically significant) for the top tercile (Table 8). This may reflect differences in test design. Since we were more concerned about test floor effects than ceiling effects, our tests focused largely on reading with comprehension at below-grade levels, while the school tests would have a much higher proportion of (more difficult) items at grade level.

[22]About 13% of lottery-winners attended for one day or less. The mean attendance among the rest was 57 days (around 66%). Figure A.2 plots the distribution of attendance among lottery winners, and Table A.4 presents correlations of attendance among lottery winners with various baseline characteristics.

actually attending the Mindspark centers, and (with further assumptions) estimates of predicted treatment effects at different levels of program exposure. We estimate the dose-response relationship between days of attendance and value-added using:

$$Y_{is2} = \alpha + \gamma.Y_{is1} + \mu_1.Attendance_i + \eta_{is2} \tag{2}$$

where Y_{ist} is defined as previously, *Attendance* is the number of days a student logged in to the Mindspark system (which is zero for all lottery-losers) and η_{ist} is the error term. Since program attendance may be endogenous to expected gains from the program, we instrument for *Attendance* with the randomized offer of a voucher.

The IV estimates suggest that, on average, an extra day of attending the Mindspark centers increased test scores by 0.0067σ in math and 0.0043σ in Hindi (Table 9: Cols. 1-2). These estimates identify the average causal response (ACR) of the treatment which "captures a weighted average of causal responses to a unit change in treatment (in this case, an extra day of attendance), for those whose treatment status is affected by the instrument" (Angrist and Imbens, 1995). Using these IV estimates to predict the effect of varying the number of days attended requires further assumptions about (a) the nature of heterogeneity in treatment effects across students (since the ACR is only identified over a subset of compliers, and not the full sample) and (b) the functional form of the relationship between days attended and the treatment effect (since the ACR averages causal effects over different intensities of treatment).

We present three pieces of suggestive evidence that constant treatment effects across students may be a reasonable assumption in this setting. First, the ITT effects were constant across the full distribution of initial achievement, which is a good summary measure for relevant individual-specific heterogeneity (Figure 3, Table 4). We also found no significant evidence of treatment heterogeneity across observed pre-treatment characteristics (Table 4).

Second, we cannot reject the equality of the IV estimates of Eq.(3) and the OLS estimates using a value-added (VA) specification (Table 9: Cols. 3-4), which suggests that the ATE and the LATE may be similar here. For both math and Hindi, the p-value from the difference-in-Sargan test (similar to a Hausman test, but allowing for heteroskedasticity) testing equivalence of OLS and IV results is substantially greater than 0.1 (Cols. 1-2).[23]

Finally, the constant term in the OLS VA specifications (corresponding to zero attendance) is similar when estimated using the full sample and when estimated using only the data in the treatment group (Table 9: Cols. 3-6).[24] The constant term is identified using both the control group and "never-takers" when using the full sample, but is identified over only

[23]Note that this close correspondence between the OLS VA and IV estimates is consistent with much recent evidence that VA models typically agree closely with experimental and quasi-experimental estimates (see, for instance Chetty, Friedman and Rockoff (2014); Deming et al. (2014); Singh (2015); Angrist et al. (2016)

[24]We cannot reject equality of the constant across regressions in either math (p=0.38) or in Hindi (p=0.61).

the "never-takers" when the sample is restricted to lottery-winners. Thus, the similarity of outcomes for the "never takers" and the control group, suggests equality of potential outcomes across different compliance groups.[25]

We next explore the functional form of the relationship between days attended and the treatment effect both graphically (by plotting value-added against attendance for the lottery winners) and analytically. The graphical analysis suggests a linear relationship in both subjects (Figure 5). Further, while test-score value added is strongly correlated with the number of days attended in a linear specification (Table 9: Cols. 3-6), adding a quadratic term does not improve fit, and the quadratic term is not significant (see Table A.5). A linear dose-response is additionally plausible when considering the adaptive nature of the intervention which allows it to be equally effective regardless of the initial learning level of the student or the rate of academic progress. Thus, diminishing returns to program exposure may not apply over the relatively short duration of treatment in this study (which is consistent with the pattern seen in Figure 5).

Under the assumptions of constant treatment effects and a linear dose-response relationship, both of which appear reasonable in this context, our IV results suggest that attending Mindspark centers for 90 days, which roughly corresponds to half a school year with 80% attendance, would lead to gains of 0.6σ in math and 0.39σ in Hindi (last row of Table 9). We extrapolate results to 90 days, rather than a full school year, to keep the predictions near the range of the program exposure provided by our experiment (the maximum was 86 days). Similar or longer durations of program exposure would be feasible, even at observed attendance rates, if for instance the intervention started at the beginning of the school year rather than midway as in this study.

These estimates are conservative and likely to understate the dose-response relationship because the *Attendance* variable includes time spent in the Mindspark centers on instruction in other subjects that we do not test (especially English).[26] In Table A.6, we present analogous IV and value-added estimates which only account for days spent by students on the subjects that we test (math and Hindi). Using these results, and the same assumptions as above, we estimate that 90 days of Mindspark attendance, split equally between the two subjects, would lead to test score gains of 0.8σ in math and 0.54σ in Hindi (last row of Table A.6).

[25]This test is similar in spirit to tests suggested by Bertanha and Imbens (2014) and Brinch, Mogstad and Wiswall (2017), for extending the validity of RD and IV estimates beyond LATE to average treatment effects.

[26]See Muralidharan and Sundararaman (2015) for an illustration of the importance of accounting for patterns of time use across subjects for inference regarding the productivity of education interventions.

4.5 Robustness

4.5.1 Attrition

Since the difference in attrition between the treatment and control groups is significant at the 10% level (Table 1), we test the robustness of our results to attrition by modeling selection into the endline based on observed characteristics, and present inverse probability weighted treatment effects: the estimated ITT effects are almost unchanged (Table A.7). We also compute Lee (2009) bounds for the ITT effect: although bounds are wide, the treatment effects are always positive and significant (Table A.8).

4.5.2 Familiarity with test questions

Our independent tests used items from several external assessments, some of which (in the Indian setting) were designed by EI; this raises the possibility that results on our assessments are overstated due to duplication of items between our tests and the Mindspark item bank. Note that this item bank contains over 45,000 items and so mere duplication in the database does not imply that a student would have been presented the same item during the intervention. Nevertheless, we test for this concern by computing the treatment effect expressed as the proportion correct on items from EI assessments and items from other assessments. The ITT effects are positive, statistically significant and of similar magnitude for both sets of items in math and Hindi (Table A.9).

4.5.3 Private Tutoring

Our results may also be confounded if winning a Mindspark voucher led to changes in the use of private tutoring. To test for this possibility, we collected data from parents of students in the experiment, using phone surveys, on whether the student attended paid extra tutoring (other than Mindspark) in any subject for each month from July 2015 to March 2016. Dividing this period into "pre-intervention" (July to September 2015) and "post-intervention" (October 2015 to March 2016), we test whether winning a Mindspark-voucher affected the incidence of private tutoring in the "post-intervention" period. We present these results in Table A.10. While there is a modest increase in private tutoring for all students in the post-treatment period (consistent with increased tutoring closer to annual school exams), we find no evidence of any differential use of private tutoring among lottery winners.

5 Discussion

5.1 Mechanisms

The estimates presented above reflect a combination of the CAL software, group teaching, and additional instructional time, and we cannot experimentally identify the relative contribution

of these channels. In this section, we present four sets of additional evidence that each point to the CAL system being the critical factor driving the large test-score gains we find.

The first, and most important, piece of evidence comes from a contemporaneous study conducted in the same location and student age group: Berry and Mukherjee (2016) report results from a randomized evaluation that studied the impact of after-school private tutoring on learning outcomes of middle-school students (in grades 6-8) in Delhi at the same time as our study. The program also provided six days of instruction per week, for three hours per day (versus 1.5 hours per day at Mindspark centers), and also charged INR 200 per month.[27] The tutoring program was run by a well-respected non-profit organization, Pratham, who have run several education programs in India that have been found to have significant positive impacts on student learning at the primary level (see, for example, Banerjee et al. (2007, 2016)). Despite several similarities, there were two key differences between this program and the Mindspark centers. First, this program focused on reinforcing knowledge of the grade-level curriculum and was not customized to students' academic preparation.[28] Second, the instruction was delivered in person by a tutor in groups of up to 20 students (a similar ratio of instructor to students as seen in Mindspark centers), but did not make use of any technology for instruction.

At the end of a year of the program, Berry and Mukherjee (2016) find *no impact* on student test scores in independent assessments of either math or language despite the program having spent more than twice the after-school instructional time provided by the Mindspark centers during our evaluation (double the scheduled instruction time per week, and evaluated after a full year as opposed to 4.5 months). These results suggest that additional instructional time with group-tutoring (the other two components of our intervention in addition to the CAL) on their own may not have had much impact on learning.[29] They also suggest that the binding constraint to student learning in this setting was not instructional time, but the (likely) ineffectiveness of additional instructional time spent on the default of teaching at a grade-appropriate level in a setting where most students are several grade levels behind (as seen in Figure 1).

Second, we provide direct evidence that the CAL software effectively addressed this constraint to effective pedagogy by targeting instructional material at the level of each individual student, and thereby accommodating the wide variation in student preparation documented in Figure 1. We see this in Figure 6, where the horizontal axis on each subgraph shows the assessed level of academic preparedness of each student enrolled in a given grade, and the vertical axis shows

[27]The average age of students in Berry and Mukherjee (2016) was 12.06 years compared to 12.67 in our study. The slight difference is due to our sample also including students in grade 9 and not just grades 6-8.

[28]While Pratham has been at the forefront of implementing the "Teaching at the Right Level (TaRL)" approach, this particular program focused on reviewing grade-level content in response to parental demand (based on personal correspondence with authors of Berry and Mukherjee (2016)).

[29]Note that these null results are unlikely to be attributable to control students attending other private tuitions instead. Berry and Mukherjee (2016) report a significant first stage on lottery winners attending *any* private tuition and can rule out effect sizes greater than 0.15σ.

that the CAL software presented students with material that is either *at their grade level or at adjacent grade levels*.[30] Further, the CAL system not only accommodates variation in initial learning levels, but also in the pace of learning across students. Figure 7 presents non-parametric plots of the average difficulty level of the math items presented to students over the course of the intervention, documenting that the software updates its estimate of student achievement levels in real time and modifies instruction accordingly. The individualization of the dynamic updating of content is highlighted further in Figure A.6 where we use student-level data to plot similar trajectories separately for *each student* in the treatment group.

Teaching effectively in a setting with such large heterogeneity in the levels and trajectories of student learning within the same grade would be very challenging even for well trained and motivated teachers. In contrast, once the CAL software is programmed to present content based on a student's assessed learning level and to adjust content at the rate of student progress, the software can handle additional heterogeneity at zero marginal cost, which is not true for a teacher.[31] Thus, the CAL software was likely to have been the key enabler for *all* students to be able to learn relative to the default of grade-appropriate pedagogy in a standard classroom setting (or in an after-school group tutoring setting).

Third, data on assignment of students into Mindspark batches (who would attend group instruction together) strongly suggests that teaching was mainly taking place on the CAL platform, with the role of the instructor being to promote adherence. We see this clearly in Figure A.5, which shows that the students in our study (who are mainly in grades 6-9), were assigned to Mindspark batches that often included students enrolled in *grades 1-5 in the same batch*. This is because EI's main consideration in assigning students to batches was the timing convenience of students and parents. Thus, EI was not concerned about having students ranging from *grades 1-9 in the same batch*, which is a classroom set up that would make very little sense for group instruction.[32]

Finally, note that the patterns of test score results we present in Section 4.3.2 are also consistent with instruction being driven mainly by the software. Gains in math test scores were seen on below grade-level questions (which is what the CAL software taught) and not on grade-level questions (which were not taught by the CAL software). This is also consistent with the

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

[31]Note that the strength of the software lies not just in its ability to personalize the level of instruction, but to do so with uniformly high-quality content at all levels (with the features described in Section 2.1). Even if a teacher wanted to review lower-grade materials in class, it would be very challenging to effectively prepare material spanning several grades and present differentiated content across students in a classroom setting.

[32]Note that prior evidence on positive impacts of group-based instruction has highlighted the importance of *homogenizing* the groups by learning level for effective instruction (Banerjee et al., 2007, 2016). Thus, it is highly unlikely that EI would have chosen to have batches that spanned so many grades unless they believed that the group instruction was second order to the instruction on the CAL system.

pattern of heterogeneity observed, both on school tests and our independent assessments, by initial learning level of students.

These four pieces of evidence all suggest that the CAL software was the key driver of the results we find. Yet, according to EI, the instructor did have an important role in promoting adherence by encouraging regular student attendance at the centers, ensuring time on task while students were in front of the computer, and supervising school homework completion and exam preparation during the group-instruction period (which parents demanded). This discussion suggests that there may be complementarities between teachers and technology. So, our results should not be interpreted as the impact of CAL software by itself, but rather as an estimate of the effect of CAL in a setting where there was also an instructor to support adherence to the CAL. Alternatively, given the null results of instructor-led after-school group tutoring found by (Berry and Mukherjee, 2016), our results can also be interpreted as showing the extent to which using technology in education can raise the productivity of an instructor.

5.2 Cost-effectiveness

Since we evaluate an after-school program, a natural comparison of cost effectiveness is with after-school private tutoring, which is widespread in our setting. The direct comparison with the results in Berry and Mukherjee (2016) suggest that after-school group-based tutoring on grade-level materials had no impact on learning in the same context even with over double the duration of exposure relative to the program we study.

A second policy-relevant comparison is with the productivity of government-run schools (from where the study subjects were recruited). Per-pupil monthly spending in these schools in Delhi was around INR 1500 (USD 22) in 2014-15; students spend 240 minutes per week on math and Hindi; and we estimate that the upper-bound of the value-added in these schools was 0.33σ in math and 0.17σ in Hindi over the 4.5-month study period. Specifically, this was the *total* value-added in the control group in Table 2, which also includes the effects of home inputs and private tutoring, and therefore likely over-estimates the value-added in public schools.

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

Steady-state costs of Mindspark at policy-relevant scales are likely to be much lower since the (high) fixed costs of product development have already been incurred. If implemented in government schools, at even a modest scale of 50 schools, per-pupil costs reduce to about USD

4 per month (including hardware costs). Above a scale of 1000 schools, the per-pupil marginal costs (software maintenance and technical support) are about USD 2 annually, which is a small fraction of the USD 150 annual cost (over 10 months) during our pilot.[33] The program thus has the potential to be very cost-effective at scale.

Further, while education spending can increase continuously over time, student time is finite. Thus, it is also useful to evaluate the effectiveness of education interventions per unit of time, *independent* of financial cost. A useful point of comparison is provided by Muralidharan (2012), who finds that providing individual-level performance bonuses to teachers in India led to test score gains of 0.54σ and 0.35σ in math and language for students exposed to the program for five years. This is one of the largest effect sizes seen to date in an experimental study on education in developing countries. Yet, we estimate that regularly attending Mindspark centers for half a year would yield similar gains (in one tenth the time).[34]

Figure 7 suggests that students who received access to the Mindspark centers improved a full grade-level in math over just 4.5 months (even with only 58% attendance). Thus, using Mindspark regularly in schools may be an especially promising option for helping to bridge the large gaps in student readiness within time frames that may make it feasible for lagging students to catch up to grade-level standards of instruction. Testing this possibility is an important topic for future research.

5.3 Policy Implications

Despite the large test-score gains we find, parental demand for Mindspark centers was low in the absence of the (fee-waiving) vouchers. In fact, all three centers in our study closed down soon after the conclusion of our experiment in the face of low parental willingness to pay (even at the subsidized price that was charged to the students outside our study who attended the Mindspark centers). The donors who subsidized the fees for regular students at Mindspark centers stipulated that they would only continue funding the subsidies if the centers could operate at or above 80% capacity (and thereby demonstrate parental willingness to pay at least the subsidized price). In practice, enrolment levels were considerably below this target, and the centers had to shut down because philanthropic funding for the subsidies ended.[35] Thus,

[33]These numbers are based on an actual budget for deploying Mindspark in government schools that was prepared and submitted by EI in 2017.

[34]Of course, it is likely that some of these gains will fade out over time as was seen in Banerjee et al. (2007). However, it is now well-known that the effects of *all* education interventions decay over time (Jacob, Lefgren and Sims, 2010; Andrabi et al., 2011). This is why we do not claim that extending the Mindspark program for 5 years will lead to ten times greater test score gains, but simply note that the gains observed over 5 years in Muralidharan (2012) were achieved in one-tenth the time here.

[35]However, Mindspark as a product is doing well and EI continues to operate and improve the full-fee Mindspark models for higher SES families, where the demand continues to be strong. Since the centers shut down in March 2016, control group students who had been offered free access to the centers after the endline test, were instead offered free educational materials as compensation for participating in the study.

models of CAL that charge fees may limit the ability of low-income students to access them and effectively deploying education technology in public schools is likely to be important for providing access to CAL programs to the most disadvantaged students.

This belief is reflected in the growing policy interest around the world in using technology in public education. However, policy makers (especially in developing countries) have mainly concentrated on providing computer hardware without commensurate attention to using technology to improve pedagogy.[36] Our results (combined with the review of evidence in Appendix C), suggest that these hardware investments are likely to yield much greater returns in terms of improved learning outcomes if attention is also paid to deploying Mindspark (or similar) software to improve pedagogy in public schools.

Our results are also relevant for policy debates on the best way to teach effectively in settings with large variation in student preparation. One widely-considered policy option is tracking of classrooms, but this may not be feasible in many developing-country settings.[37] Further, even when feasible, tracking is controversial and the global evidence on its impact is mixed (Betts, 2011). Our results suggest that well-designed CAL programs may be able to deliver the pedagogical advantages of tracking while mitigating several limitations, as listed below.

First, CAL allows instruction to be individualized at the student level, whereas tracked classrooms still have to cater to variation in student learning levels and trajectories with a common instruction protocol. Second, by allowing students to work at their own pace, it avoids potential negative effects of students being labelled as being in a weaker track. Third, the dynamic updating of content mitigates the risk of premature permanent tracking of 'late bloomers'. Fourth, it allows instruction to be differentiated without changing peers in the classroom. Fifth, relative to policies of grade retention or accelerated grade promotion, using CAL programs in classrooms makes it possible to preserve the age-cohort based social grouping of students (which may allow for better development of socio-emotional skills), while allowing for variation in academic content presented.

6 Conclusion

We present experimental evidence on the impact of a technology-led supplementary instruction program in post-primary grades in urban India, and find that gaining access to the program led to large and rapid test-score gains in both math and language. The combination of facts

[36]For instance, various state governments in India have distributed free laptops to students in recent years. Further, progress on implementing the national-level policy on technology in education is typically measured by the number of schools with computer labs.

[37]Unlike in developed countries where students in middle and high schools can choose their subjects and can take easier and more advanced courses, most developing-country education systems in South Asia and sub-Saharan Africa are characterized by preparing students for a single high-stakes school leaving examination. Thus, the default organization of schools is to have all students in a given grade in the same classroom with the teacher focusing on completing the curriculum mandated by official text books for the corresponding grade.

presented in Figures 1 and 6 highlight both the challenge of effective teaching in conditions with large levels of heterogeneity in student learning, and the promise of computer-aided learning (CAL) to address this challenge by being able to "Teach at the Right Level" (TaRL) for all students. We therefore conjecture that a key reason for the large effects we find is the ability of the CAL program to teach *all* students equally effectively including those left behind by business-as-usual instruction (as seen in Figure 4).

In addition to effectively implementing TaRL, the large effects may also reflect the software's ability to effectively address other constraints to effective teaching and learning. The high quality of content, combined with effective delivery and interface, may help circumvent constraints of teacher human capital and motivation. The structure of the content (requiring regular student interaction with the system) may also help to promote student engagement relative to passive participation in typical classroom instruction. Algorithms for analyzing patterns of student errors and providing differentiated feedback and follow up content that is administered in real-time, allows for feedback that is more relevant and much more frequent. These features all reflect continuous and iterative program development over a long period of more than a decade.

These effects may plausibly be increased even further with better design. It is possible that in-school settings may have greater adherence to the program in terms of attendance. It may also be possible to improve the effectiveness of teacher-led instruction in a 'blended learning' environment by using the extensive information on student-performance to better guide teacher effort in the classroom. This 'big data' on student achievement also offers much potential of its own. In particular, such a setting may enable high-frequency randomized experiments on effective pedagogical techniques and approaches (which may vary across students) and build a stronger evidence base on effective teaching practices. This evidence may then be used to further optimize the delivery of instruction in the program and, plausibly, also for the delivery of classroom instruction. Finally, the detailed and continuous measures of effort input by the students can be used directly to reward students, with potentially large gains in student motivation, effort, and achievement.[38]

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

[38]Direct evidence that this may be possible is provided by Hirshleifer (2015) who uses data from a (different) computer-aided instruction intervention to reward student effort and documents large effects of 0.57σ.

away current instructional time; and whether it is delivered without supervision, under the supervision of current teachers, or under the supervision of third parties (e.g. the Mindspark center staff). Identifying the most effective modes of delivery for the program at larger scale is an important area for future research.[39]

A further point of caution is that our results should not be interpreted as supporting a de-emphasis of the role of teachers in education. Rather, since the delivery of education involves several non-routine tasks that vary as a function of individual students and situations, and requires complex contextually-aware communication, it is more likely that technology will complement rather than substitute teachers (as shown more generally by Autor, Levy and Murnane (2003)). So, it may be possible to improve teacher and school productivity by using technology to perform routine tasks (such as grading) and data-analysis intensive tasks (such as identifying patterns in student answers and providing differentiated feedback and instruction to students), and enabling teachers to spend more time on aspects of education where they may have a comparative advantage - such as supporting group-based learning strategies that may help build social and other non-cognitive skills that may have considerable labor market returns (Cunha, Heckman and Schennach, 2010; Heckman and Kautz, 2012; Deming, 2017).

Overall, our study is best regarded as an efficacy trial documenting that well-designed and implemented technology-enabled learning programs can produce large gains in student test scores in a relatively short period of time. The promise of such an approach may be especially high in developing country settings that feature large levels of heterogeneity in student learning levels across students enrolled in the same grade, and a default of textbook- and curriculum-based instruction that leaves many students behind (as seen in our data). There is robust evidence across settings that pedagogical approaches that enable "Teaching at the Right Level" (TaRL) are highly effective, but it is non-trivial to scale these up. Our results suggest that the promise of technology to implement TaRL and sharply improve productivity in the delivery of education is real, and that there may be large returns to further innovation and research on effective ways of integrating technology-aided instruction into classrooms, and on effective ways of delivering these benefits at a larger scale.

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

References

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

Angrist, Joshua, and Guido 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.

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

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2016. "Leveraging lotteries for school value-added: Testing and estimation." *The Quarterly Journal of Economics*, Forthcoming.

Autor, David, Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics*, 118(4): 1279–1333.

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

Banerjee, Abhijit, Paul Glewwe, Shawn Powers, and Melanie Wasserman. 2013. "Expanding access and increasing student learning in post-primary education in developing countries: A review of the evidence." Abdul Latif Jameel Poverty Action Lab.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of ?Teaching at the Right Level? in India." National Bureau of Economic Research, Inc NBER Working Papers 22746.

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

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

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

Bertanha, Marinho, and Guido Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." National Bureau of Economic Research, Inc 20773.

Betts, Julian. 2011. "The Economics of Tracking in Education." In *Handbook of the Economics of Education.* , ed. Eric Hanushek, Stephen Machin and Ludger Woessmann, 341–381. Elsevier.

Beuermann, Diether W, Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu 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.

Bhattacharjea, S., W. Wadhwa, and R. Banerji. 2011. *Inside primary schools: A study of teaching and learning in rural India.* ASER Centre, New Delhi.

Bold, Tessa, Deon P. Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian William Stacy, Jakob Svensson, and Waly Wane. 2017. "What do teachers know and do? Does it matter? Evidence from primary schools in Africa." The World Bank Policy Research Working Paper Series 7956.

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.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125(4): 985–1039.

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

Buswell, Guy Thomas, and Charles Hubbard Judd. 1925. *Summary of educational investigations relating to arithmetic.* University of Chicago.

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, Paul E, Mercedes Onofa, and Juan 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, Raj, John N Friedman, and Jonah 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, Jessica, and Pascaline Dupas. 2010. "Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment." *The Quarterly Journal of Economics*, 125(1): 1–45.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio 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.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica*, 78(3): 883–931.

Das, Jishnu, and Tristan 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.

Deming, David J. 2017. "The growing importance of social skills in the labor market." *The Quarterly Journal of Economics*, 132(4): 1593–1640.

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

Dewan, Hridaykant, Namrita Batra, and Inder Singh Chabra. 2012. "Transforming the Elementary Mathematics Curriculum: Issues and Challenges." In *Mathematics Education in India: Status and Outlook*. , ed. R. Ramanujan and K. Subramaniam. Mumbai, India:Homi Bhabha Centre for Science Education, Tata Institute for Fundamental Research.

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

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

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education*. , ed. Eric Hanushek, Stephen Machin and Ludger Woessmann, 653–744. Elsevier.

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

Heckman, James J., and Tim Kautz. 2012. "The Economics of Human Development and Social Mobility." *Labour Economics*, 19(4): 451–464.

Hirshleifer, Sarojini. 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, Brian A, Lars Lefgren, and David P Sims. 2010. "The persistence of teacher-induced learning." *Journal of Human resources*, 45(4): 915–943.

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

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

Kumar, Ruchi S., Hridaykant Dewan, and K.Subramaniam. 2012. "The preparation and professional development of mathematics teachers." In *Mathematics Education in India: Status and Outlook*. , ed. R. Ramanujan and K. Subramaniam. Mumbai, India:Homi Bhabha Centre for Science Education, Tata Institute for Fundamental Research.

Lai, Fang, Linxiu Zhang, Qinghe Qu, Xiao Hu, Yaojiang Shi, Matthew Boswell, and Scott 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.

Lai, Fang, Linxiu Zhang, Xiao Hu, Qinghe Qu, Yaojiang Shi, Yajie Qiao, Matthew Boswell, and Scott 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, Fang, Renfu Luo, Linxiu Zhang, and Scott 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.

Lee, David. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies*, 76: 1071–1102.

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek, and Dinand 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, Stephen, Sandra McNally, and Olmo Silva. 2007. "New technology in schools: Is there a payoff?" *The Economic Journal*, 117(522): 1145–1167.

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

Mo, Di, Johan Swinnen, Linxiu Zhang, Hongmei Yi, Qinghe Qu, Matthew Boswell, and Scott 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, Di, Linxiu Zhang, Renfu Luo, Qinghe Qu, Weiming Huang, Jiafu Wang, Yajie Qiao, Matthew Boswell, and Scott Rozelle. 2014*a*. "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, Di, L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell, and S. Rozelle. 2014*b*. "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).

Mo, Di, Yu Bai, Matthew Boswell, and Scott Rozelle. 2016. "Evaluating the effectiveness of computers as tutors in China."

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.

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

Muralidharan, Karthik. 2017. "Field Experiments in Education in Developing Countries." In *Handbook of Field Experiments*. , ed. Abhijit Banerjee and Esther Duflo. Elsevier.

Muralidharan, Karthik, and Abhijeet Singh. 2018. "Improving Public Sector Governance at Scale: Experimental Evidence from a Large-Scale School Governance Improvement Program in India." University of California San Diego mimeo., San Diego, CA.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. "The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India." *The Economic Journal*, 120(F187-F203).

Muralidharan, Karthik, and Venkatesh 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.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017. "The fiscal cost of weak governance: Evidence from teacher absence in India." *Journal of Public Economics*, 145: 116–135.

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

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.

NCERT. 2006. *Position Paper of the National Focus Group on Curriculum, Syllabus and Textbooks.* National Council of Educational Research and Training, New Delhi.

PASEC. 2015. *Programme d?Analyse des Systmes ducatifs de la Confemen (PASEC) 2014: Education System Performance in Francophone Africa, Competencies and Learning Factors in Primary Education.* PASEC, Dakar, Senegal.

Pearson, P.D., R.E. Ferdig, R.L. 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.

Pratham. 2016. *Annual Status of Education Report 2015.* Pratham, New Delhi.

Pratham. 2017. *Annual Status of Education Report 2016.* Pratham, New Delhi.

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

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

Rampal, Anita, and Jayasree Subramaniam. 2012. "Transforming the Elementary Mathematics Curriculum: Issues and Challenges." In *Mathematics Education in India: Status and Outlook.* , ed. R. Ramanujan and K. Subramaniam. Mumbai, India:Homi Bhabha Centre for Science Education, Tata Institute for Fundamental Research.

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

Rouse, Cecilia Elena, and Alan 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.

SACMEQ. 2007. *Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), Various years.* University of Botswana, Gaborone. http://www.sacmeq.org/.

SAFED. 2017. *Annual Status of Education Report (ASER-Pakistan) 2016.* South Asia Forum for Education Development, Lahore.

Sankar, Deepa, and Toby Linden. 2014. "How much and what kind of teaching is there in elementary education in India? Evidence from three states." (South Asia Human Development Sector Report No. 67). Washington, DC: The World Bank.

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

Sinha, S., R. Banerji, and W. Wadhwa. 2016. *Teacher performance in Bihar, India: Implications for education.* The World Bank, Washington D.C.

Uwezo. 2016. *Are Our Children Learning? Uwezo Uganda 6th Learning Assessment Report.* Twaweza East Africa, Kampala.

van der Linden, Wim J, and Ronald K Hambleton. 2013. *Handbook of modern item response theory.* Springer Science & Business Media.

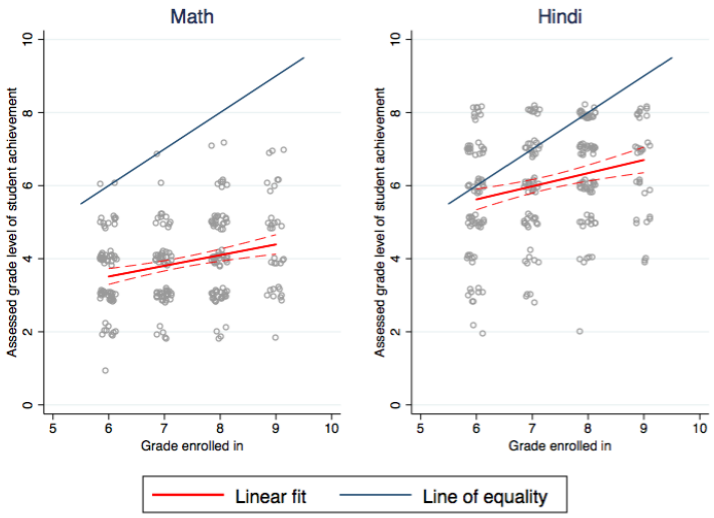
Waxman, H.C., M.-F. Lin, and G.M. Michko. 2003. "A meta-analysis of the effectiveness of teaching and learning with technology on student outcomes." *Unpublished manuscript.* CambridgeNaperville, IL: Learning Point Associates.

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

World Bank. 2016. *What is happening inside classrooms in Indian secondary schools? A time on task study in Madhya Pradesh and Tamil Nadu.* The World Bank, Washington D.C.

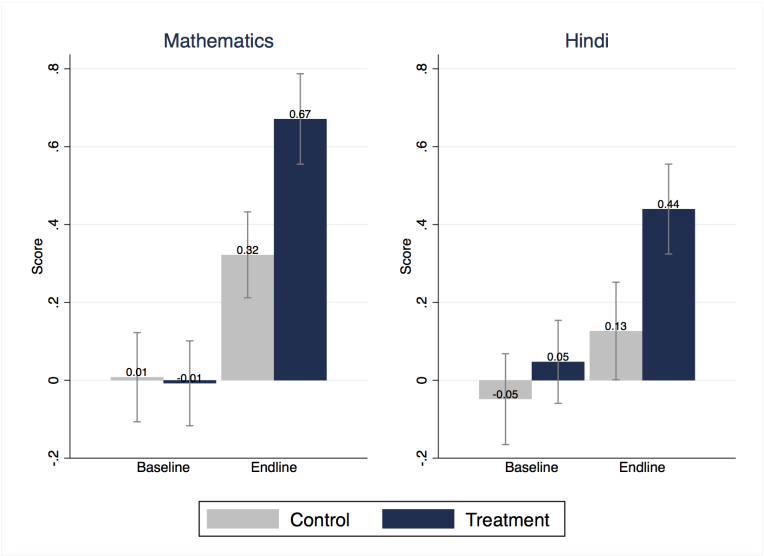
World Bank. 2018. *World Development Report 2018: Learning to Realize Education?s Promise.* World Bank, Washington, DC.

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



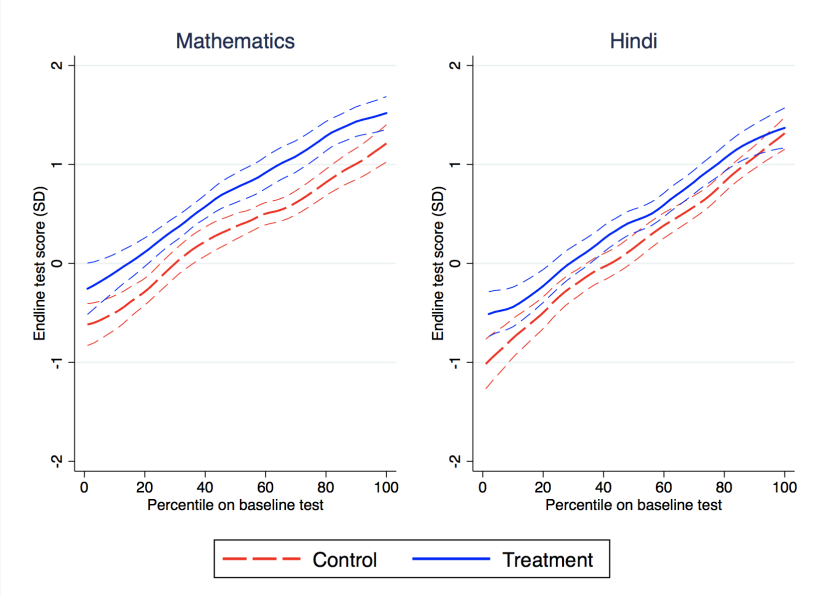
Note: This figure shows, for treatment group, the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. These data are from the *initial* diagnostic test, and do not reflect any instruction provided by Mindspark. In both subjects, we find 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: Mean difference in test scores between lottery winners and losers



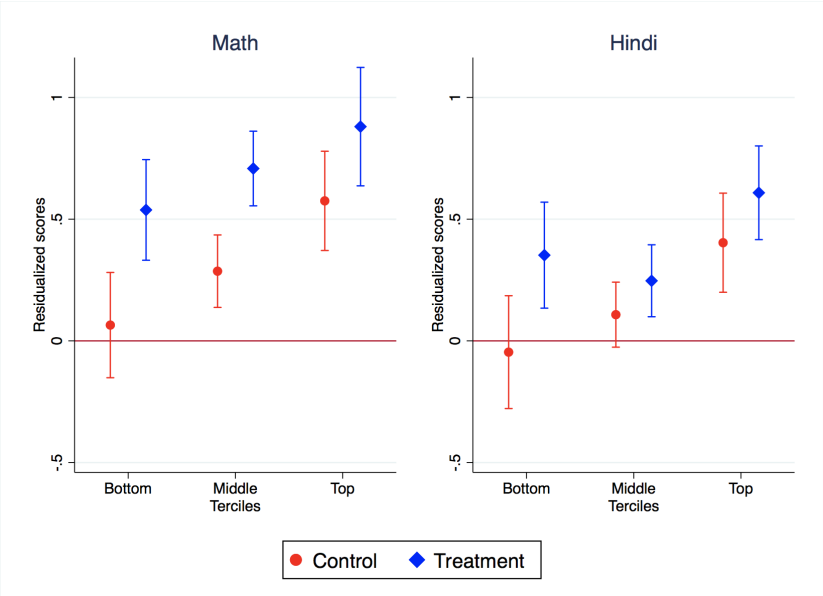
Note: This figure shows mean of test scores, normalized with reference to baseline, across treatment and control groups in the two rounds of testing with 95% confidence intervals. Test scores were linked within-subject through IRT models, pooling across grades and across baseline and endline, and are 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 3: Non-parametric investigation of treatment effects by baseline percentiles



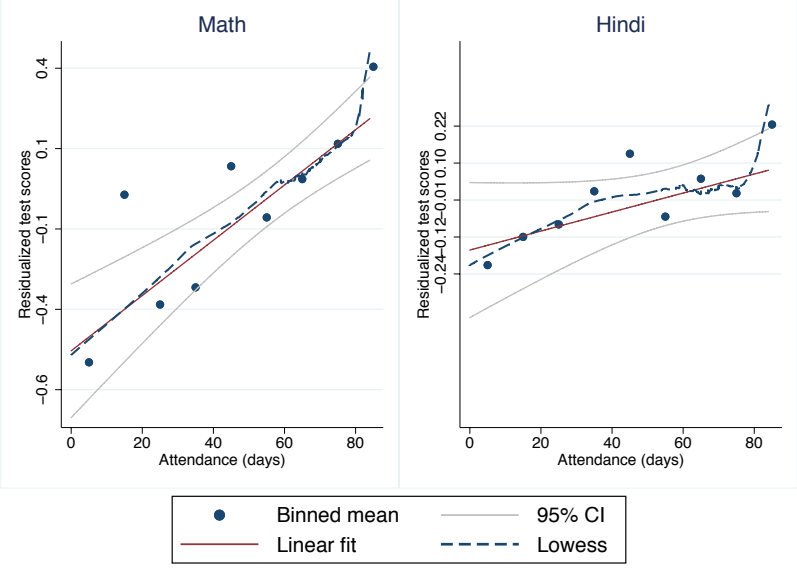
Note: The figures present kernel-weighted local mean smoothed plots which relate endline test scores to percentiles in the baseline achievement, separately for the treatment and control groups, alongside 95% confidence intervals. At all percentiles of baseline achievement, treatment group students score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

Figure 4: Growth in achievement in treatment and control groups



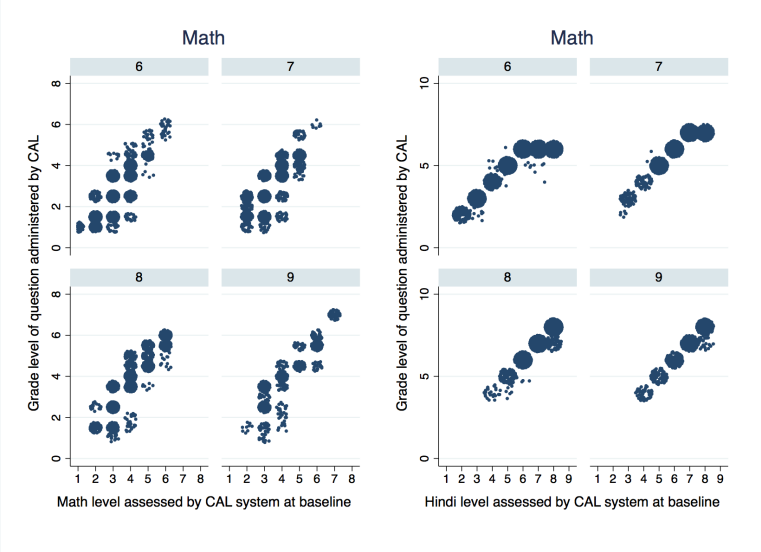
Note: This figure shows the growth in student achievement in the treatment and control groups in math and Hindi, as in Table 5. Students in the treatment group see positive value-added in all terciles whereas we cannot reject the null of no academic progress for students in the bottom tercile in the control group.

Figure 5: Dose response relationship



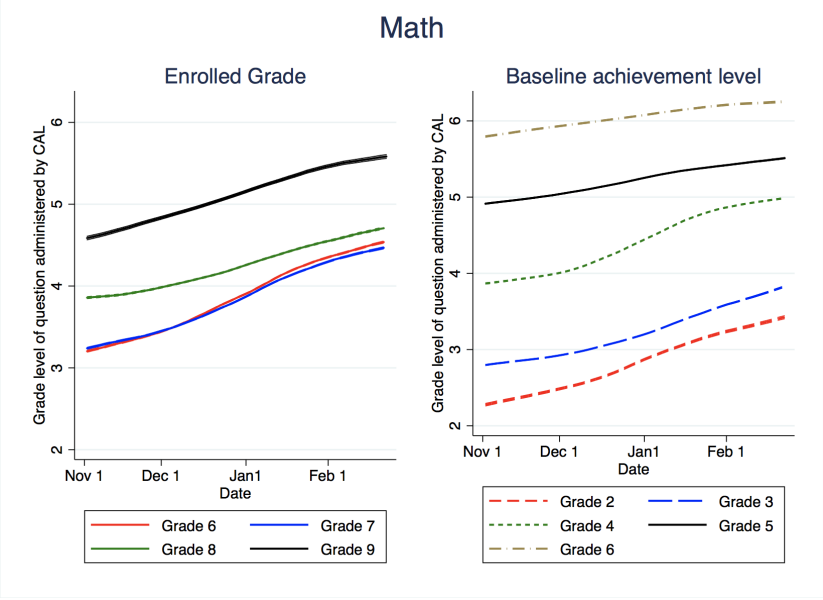
Note: This figure explores the relationship between value-added and attendance in the Mindspark program among the lottery-winners. It presents the mean value-added in bins of attendance along with a linear fit and a lowess smoothed non-parametric plot.

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



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

Figure 7: Dynamic updating and individualization of content in Mindspark



Note: This figure shows kernel-weighted local mean smoothed lines relating the level of difficulty of the math questions administered to students in the treatment group with the date of administration. The left panel presents separate lines by the actual grade of enrolment. The right panel presents separate lines by the level of achievement assessed at baseline by the CAL software. Note that 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.

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	0.004	0.034	314	305
Age (years)	12.67	12.41	0.267	0.143	230	231
SES index	-0.03	0.04	-0.070	0.137	314	305
Grade in school						
Grade 4	0.01	0.01	-0.003	0.007	305	299
Grade 5	0.01	0.02	-0.007	0.010	305	299
Grade 6	0.27	0.30	-0.035	0.037	305	299
Grade 7	0.26	0.26	0.005	0.036	305	299
Grade 8	0.30	0.28	0.017	0.037	305	299
Grade 9	0.15	0.13	0.024	0.028	305	299
Baseline test scores						
Math	-0.01	0.01	-0.016	0.081	313	304
Hindi	0.05	-0.05	0.096	0.080	312	305
Present at endline	0.85	0.90	-0.048	0.027	314	305
		Panel B: Only students present in Endline				
Demographic characteristics						
Female	0.77	0.76	0.013	0.036	266	273
Age (years)	12.61	12.37	0.243	0.156	196	203
SES index	-0.17	0.03	-0.193	0.142	266	273
Grade in school						
Grade 4	0.01	0.01	-0.003	0.008	258	269
Grade 5	0.01	0.02	-0.011	0.011	258	269
Grade 6	0.28	0.30	-0.022	0.040	258	269
Grade 7	0.26	0.26	-0.001	0.038	258	269
Grade 8	0.30	0.28	0.020	0.040	258	269
Grade 9	0.14	0.12	0.017	0.029	258	269
Baseline test scores						
Math	-0.03	-0.00	-0.031	0.086	265	272
Hindi	0.05	-0.07	0.124	0.084	266	273

Note: Treatment and control here refer to groups who were randomly assigned to receive an offer of Mindspark voucher 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 are taken and (b) assessment of skills in math and Hindi, administered using pen-and-paper tests. Tests were designed to cover wide ranges of achievement and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline. The SES index refers to a wealth index generated using the first factor from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household.

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

	(1)	(2)	(3)	(4)
	Dep var: Standardized IRT scores (endline)			
	Math	Hindi	Math	Hindi
Treatment	0.37	0.23	0.37	0.24
	(0.064)	(0.062)	(0.064)	(0.071)
Baseline score	0.58	0.71	0.57	0.68
	(0.042)	(0.040)	(0.051)	(0.033)
Constant	0.33	0.17	0.32	0.17
	(0.044)	(0.044)	(0.031)	(0.035)
Strata fixed effects	Y	Y	N	N
Observations	535	537	535	537
R-squared	0.403	0.493	0.397	0.473

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

Table 3: 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	0.072	0.042	0.071	0.15	0.15	0.11
	(0.016)	(0.016)	(0.021)	(0.020)	(0.024)	(0.022)	(0.028)
Baseline math score	0.13	0.11	0.082	0.093	0.052	0.068	0.099
	(0.0080)	(0.010)	(0.015)	(0.012)	(0.014)	(0.012)	(0.016)
Constant	0.66	0.50	0.38	0.33	0.39	0.45	0.36
	(0.0079)	(0.0076)	(0.010)	(0.010)	(0.012)	(0.011)	(0.014)
Observations	537	537	537	537	537	537	537
R-squared	0.357	0.229	0.097	0.157	0.097	0.135	0.112

(b) Hindi

	(1)	(2)	(3)	(4)
	Dep var: Proportion of questions answered correctly			
	Sentence completion	Retrieve explicitly stated information	Make straightforward inferences	Interpret and integrate ideas and information
Treatment	0.046	0.045	0.065	0.053
	(0.022)	(0.016)	(0.022)	(0.015)
Baseline Hindi score	0.13	0.14	0.15	0.067
	(0.017)	(0.0075)	(0.011)	(0.013)
Constant	0.72	0.59	0.51	0.31
	(0.011)	(0.0078)	(0.011)	(0.0077)
Observations	539	539	539	539
R-squared	0.182	0.380	0.309	0.136

Note: Robust standard errors in parentheses. 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. All test questions were multiple choice items with four choices. 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 voucher till March 2016. All regressions include randomization strata fixed effects.

Table 4: Heterogeneity in treatment effect by gender, socio-economic status and baseline score

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Standardized IRT scores (endline)					
COVARIATES	Female		SES		Baseline score	
	Math	Hindi	Math	Hindi	Math	Hindi
Treatment	0.47	0.27	0.38	0.26	0.37	0.24
	(0.14)	(0.095)	(0.065)	(0.062)	(0.064)	(0.070)
Covariate	-0.050	0.21	-0.0028	0.099	0.53	0.70
	(0.14)	(0.15)	(0.035)	(0.021)	(0.076)	(0.047)
Interaction	**-0.13**	**-0.046**	**0.023**	**-0.0041**	**0.081**	**-0.047**
	(0.14)	**(0.12)**	**(0.050)**	**(0.041)**	**(0.087)**	**(0.071)**
Observations	535	537	535	537	535	537
R-squared	0.399	0.474	0.398	0.494	0.399	0.473

Note: Robust standard errors in parentheses. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark voucher till March 2016. The SES index and test scores are defined as in Tables 1 and 2 respectively. All regressions include strata fixed effects and control for baseline subject scores.

Table 5: Heterogeneity in treatment effect by within-grade terciles

	(1)	(2)
	Dep var: Standardized IRT scores (endline)	
VARIABLES	Math	Hindi
Bottom Tercile	0.13	-0.072
	(0.098)	(0.10)
Middle Tercile	0.30	0.14
	(0.073)	(0.068)
Top Tercile	0.53	0.46
	(0.092)	(0.085)
Treatment	0.33	0.41
	(0.12)	(0.12)
Treatment*Middle Tercile	0.083	-0.30
	(0.16)	(0.16)
Treatment*Top Tercile	0.068	-0.24
	(0.16)	(0.15)
Baseline test score	0.44	0.58
	(0.066)	(0.062)
Observations	535	537
R-squared	0.545	0.545

Note: Robust standard errors in parentheses.Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark voucher till March 2016. Test scores are scaled as in Table 2.

Table 6: 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	0.0089	0.081	0.063	0.050
	(0.032)	(0.013)	(0.027)	(0.014)
Baseline subject score	0.047	0.099	0.13	0.13
	(0.022)	(0.0069)	(0.016)	(0.0068)
Constant	0.31	0.49	0.45	0.58
	(0.022)	(0.0089)	(0.019)	(0.0100)
Observations	291	511	292	513
R-squared	0.029	0.346	0.250	0.399

Note: Robust standard errors in parentheses. The table shows the impact of the treatment (winning a randomly-assigned voucher) on questions below or at/above grade levels for individual students. The dependent variable is the proportion of questions that a student answered correctly. All test questions were multiple choice items with four choices. Our endline assessments, designed to be informative at students' actual levels of achievement, did not include many items at grade 8 level and above. Therefore students in Grades 8 and 9 are not included in regressions on items at/above grade level. Baseline scores are IRT scores in the relevant subject from the baseline assessment. All regressions include randomization strata fixed effects.

Table 7: 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	0.196	0.059	0.077	0.108	0.081	0.100
	(0.088)	(0.076)	(0.092)	(0.110)	(0.105)	(0.080)
Baseline Hindi score	0.487		0.292	0.414	0.305	0.336
	(0.092)		(0.064)	(0.096)	(0.067)	(0.058)
Baseline math score		0.303	0.097	0.262	0.120	0.167
		(0.041)	(0.036)	(0.058)	(0.052)	(0.039)
Constant	1.006	0.142	0.931	1.062	1.487	0.977
	(1.103)	(0.423)	(0.347)	(0.724)	(0.740)	(0.600)
Observations	597	596	595	594	597	597
R-squared	0.190	0.073	0.121	0.177	0.144	0.210

Note: Robust standard errors in parentheses. 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. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark voucher till March 2016. Test scores in the school exams are normalized within school*grade to have a mean of zero and a standard deviation of one in the control group. All regressions include grade and school fixed effects.

Table 8: Heterogeneous effects on school tests, by terciles of baseline achievement

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: School test scores					
VARIABLES	Hindi	Math	Science	Soc. Sc.	English	Aggregate
Treatment	0.058	-0.40	-0.15	-0.17	0.14	-0.052
	(0.14)	(0.11)	(0.16)	(0.16)	(0.11)	(0.099)
Treatment*Tercile 2	0.11	0.55	0.31	0.15	-0.30	0.063
	(0.23)	(0.20)	(0.18)	(0.24)	(0.14)	(0.16)
Treatment*Tercile 3	0.29	0.82	0.36	0.65	0.14	0.38
	(0.18)	(0.27)	(0.19)	(0.24)	(0.15)	(0.13)
Tercile 2	-0.35	-0.27	-0.39	-0.61	0.14	-0.29
	(0.27)	(0.23)	(0.18)	(0.29)	(0.17)	(0.19)
Tercile 3	-0.23	-0.48	-0.32	-1.02	0.096	-0.37
	(0.31)	(0.21)	(0.21)	(0.38)	(0.20)	(0.21)
Baseline Hindi score	0.53		0.35	0.67	0.25	0.40
	(0.17)		(0.083)	(0.19)	(0.11)	(0.10)
Baseline Math score		0.33	0.096	0.27	0.11	0.16
		(0.072)	(0.033)	(0.058)	(0.051)	(0.039)
Constant	1.28	0.47	1.27	1.76	1.29	1.24
	(1.09)	(0.40)	(0.39)	(0.76)	(0.74)	(0.60)
Observations	597	596	595	594	597	597
R-squared	0.201	0.098	0.132	0.203	0.155	0.226
	Treatment Effect by tercile (p-values in brackets)					
Tercile 1	0.058	-0.40	-0.15	-0.17	0.14	-0.052
	[0.67]	[0.002]	[0.36]	[0.31]	[0.23]	[0.61]
Tercile 2	0.17	0.15	0.16	-0.02	-0.16	0.01
	[0.27]	[0.28]	[0.13]	[0.94]	[0.25]	[0.92]
Tercile 3	0.348	0.42	0.21	0.48	0.28	0.33
	[0.04]	[0.07]	[0.16]	[0.04]	[0.08]	[0.03]

Note: Robust standard errors in parentheses. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark voucher till March 2016. Test scores are scaled as in Table 7.

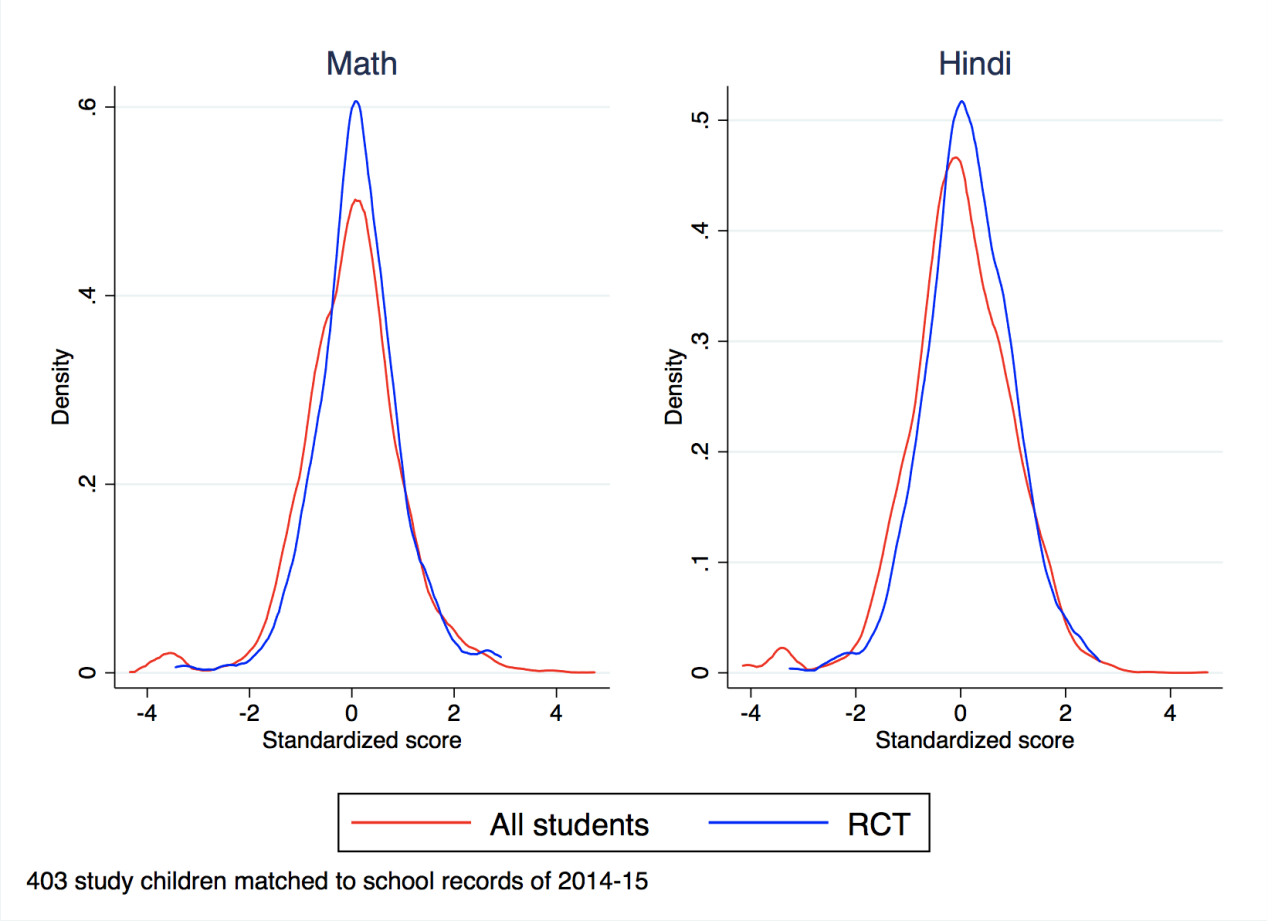
Table 9: Dose-response of Mindspark attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Standardized IRT scores (endline)					
	IV estimates		OLS VA (full sample)		OLS VA (Treatment group)	
VARIABLES	Math	Hindi	Math	Hindi	Math	Hindi
Attendance (days)	0.0067	0.0043	0.0072	0.0037	0.0086	0.0030
	(0.0011)	(0.0011)	(0.00090)	(0.00091)	(0.0018)	(0.0018)
Baseline score	0.56	0.68	0.58	0.71	0.62	0.68
	(0.038)	(0.036)	(0.042)	(0.040)	(0.061)	(0.052)
Constant			0.31	0.18	0.22	0.24
			(0.041)	(0.041)	(0.12)	(0.11)
Observations	535	537	535	537	264	265
R-squared	0.431	0.479	0.429	0.495	0.446	0.445
Angrist-Pischke F-statistic for weak instrument	1207	1244				
Diff-in-Sargan statistic for exogeneity (p-value)	0.14	0.92				
Extrapolated estimates of 90 days' treatment (SD)	0.603	0.39	0.648	0.333	0.77	0.27

Note: Robust standard errors in parentheses. Treatment group students who were randomly-selected for the Mindspark voucher 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) instrument attendance in Mindspark with the randomized allocation of a scholarship and include randomization strata fixed effects, Columns (3) and (4) present OLS value-added models in the full sample, Columns (5) and (6) present OLS value-added models using only data on the lottery-winners. Scores are scaled here as in Table 2.

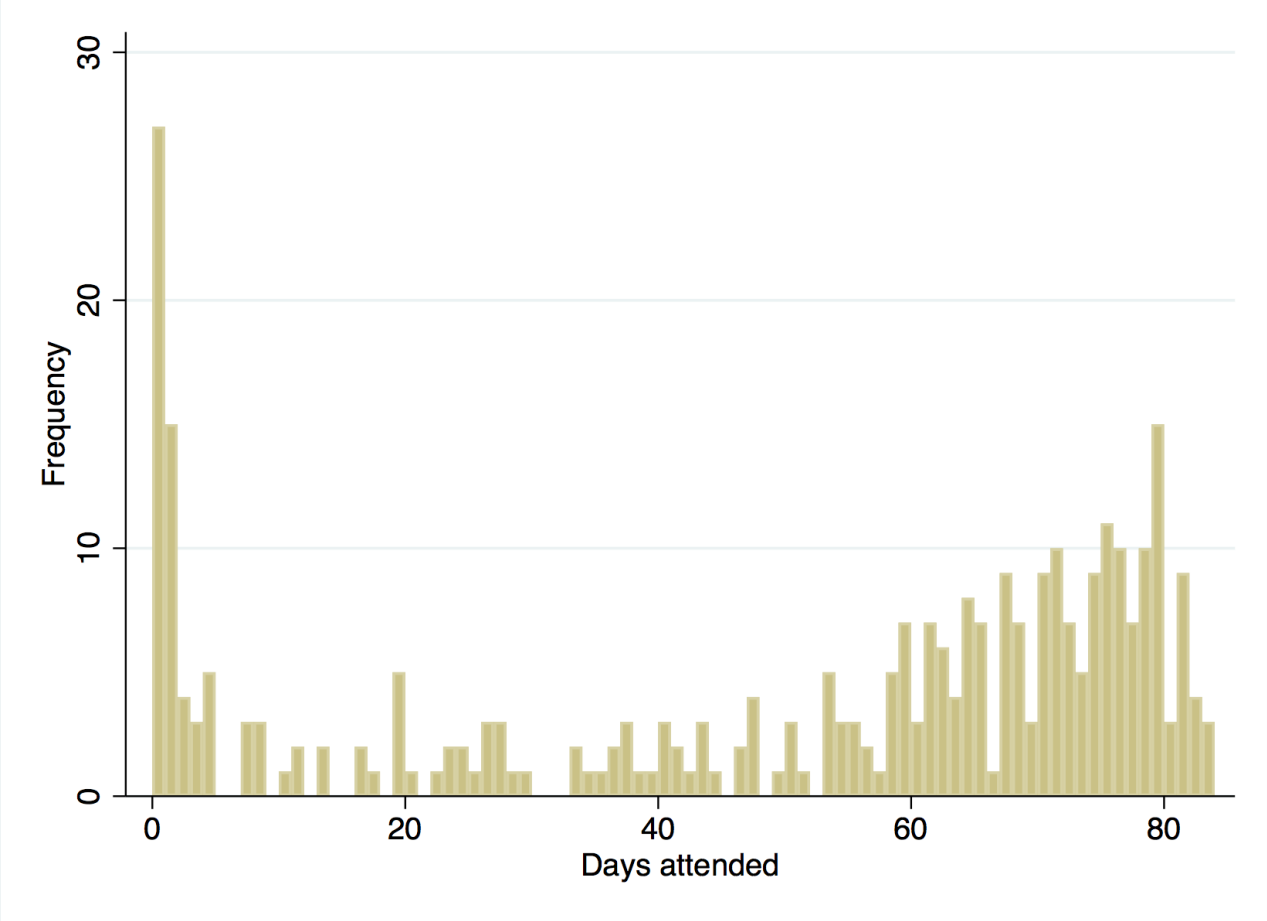
Appendix A Additional figures and tables

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



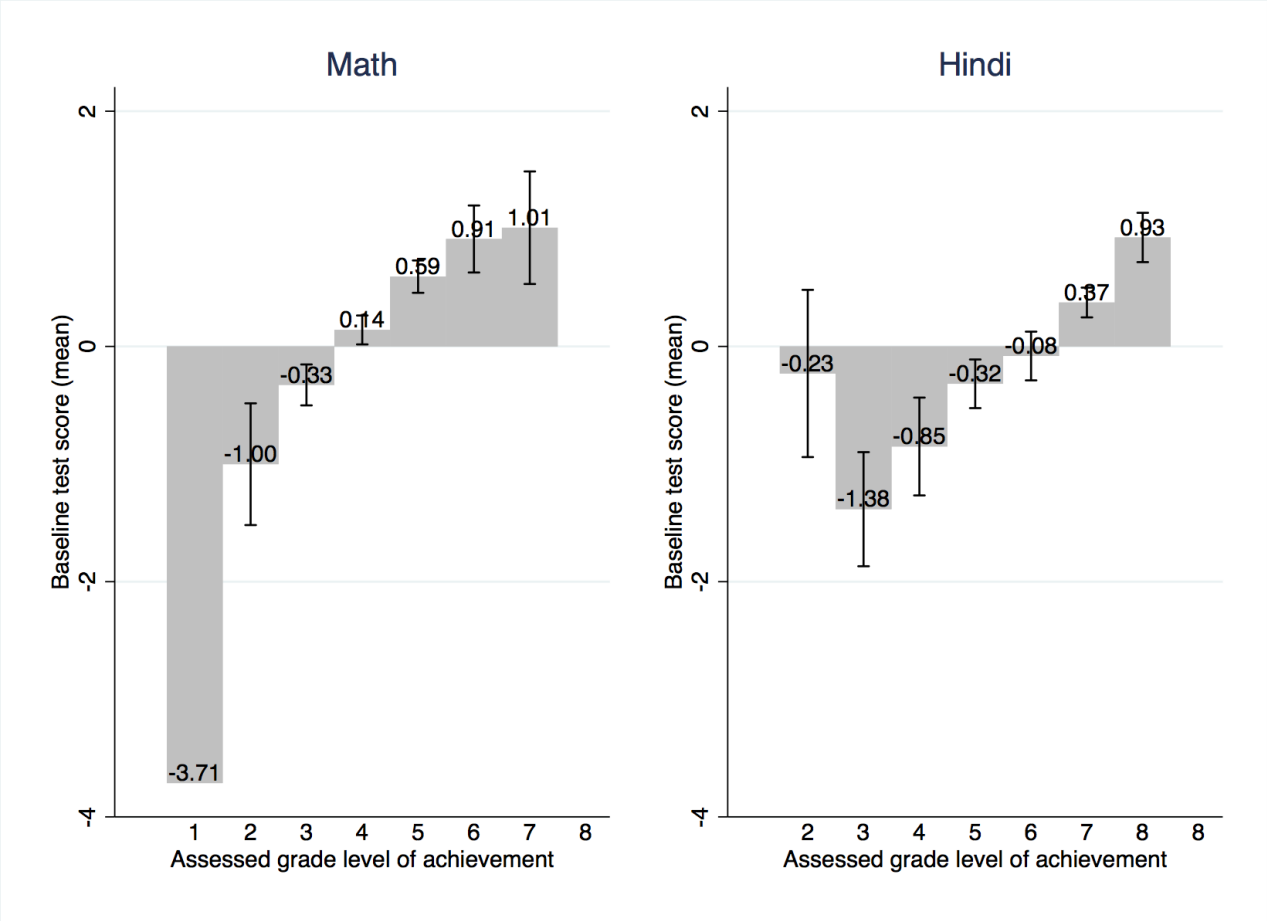
Note: The panels compare the final scores for the 2014-15 school year, i.e. the pre-program academic year, for study participants and non-participants. Test scores have been standardized within school*grade cells. The study participants are positively selected into the RCT in comparison to their peers but the magnitude of selection is modest and there is near-complete common support between the two groups in pre-program academic achievement. See Table A.1 for further details.

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



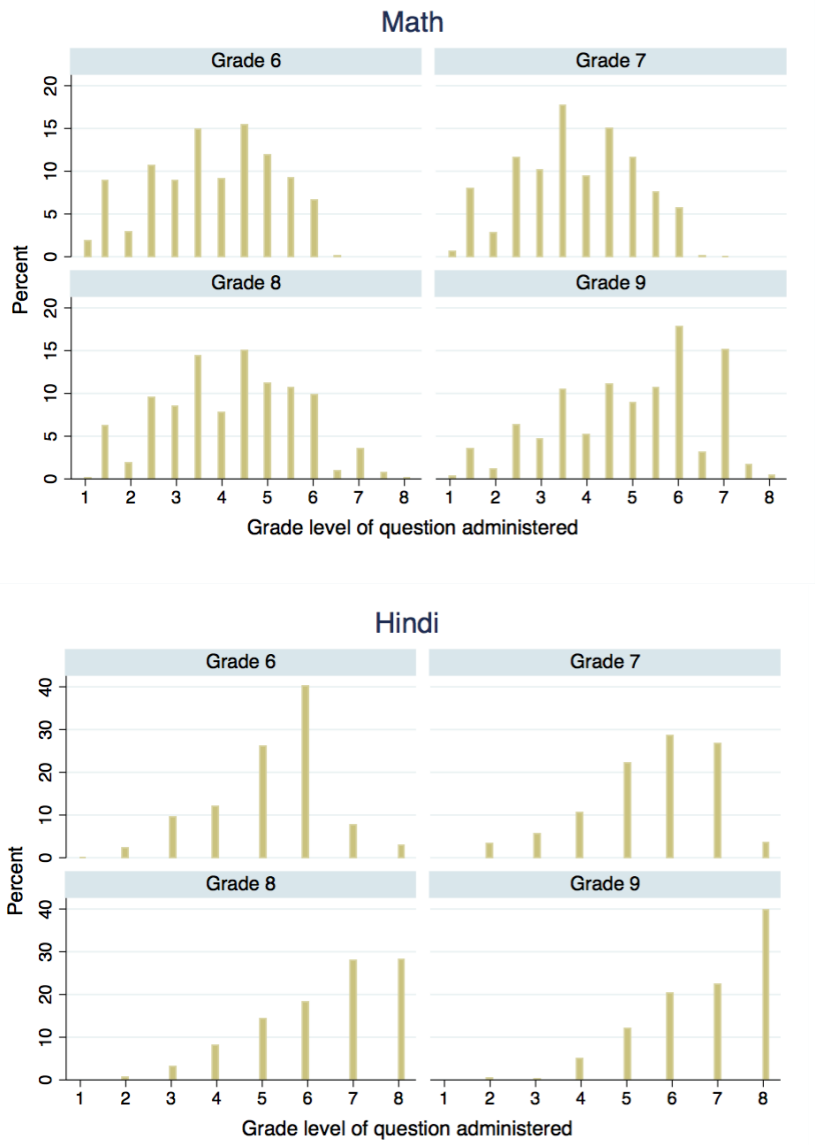
Note: This figure shows the distribution of attendance in the Mindspark centers among the lottery-winners. Over the study period, the Mindspark centers were open for 86 working days.

Figure A.3: Comparison of Mindspark initial assessment of grade-level of student achievement with (independent) baseline test scores



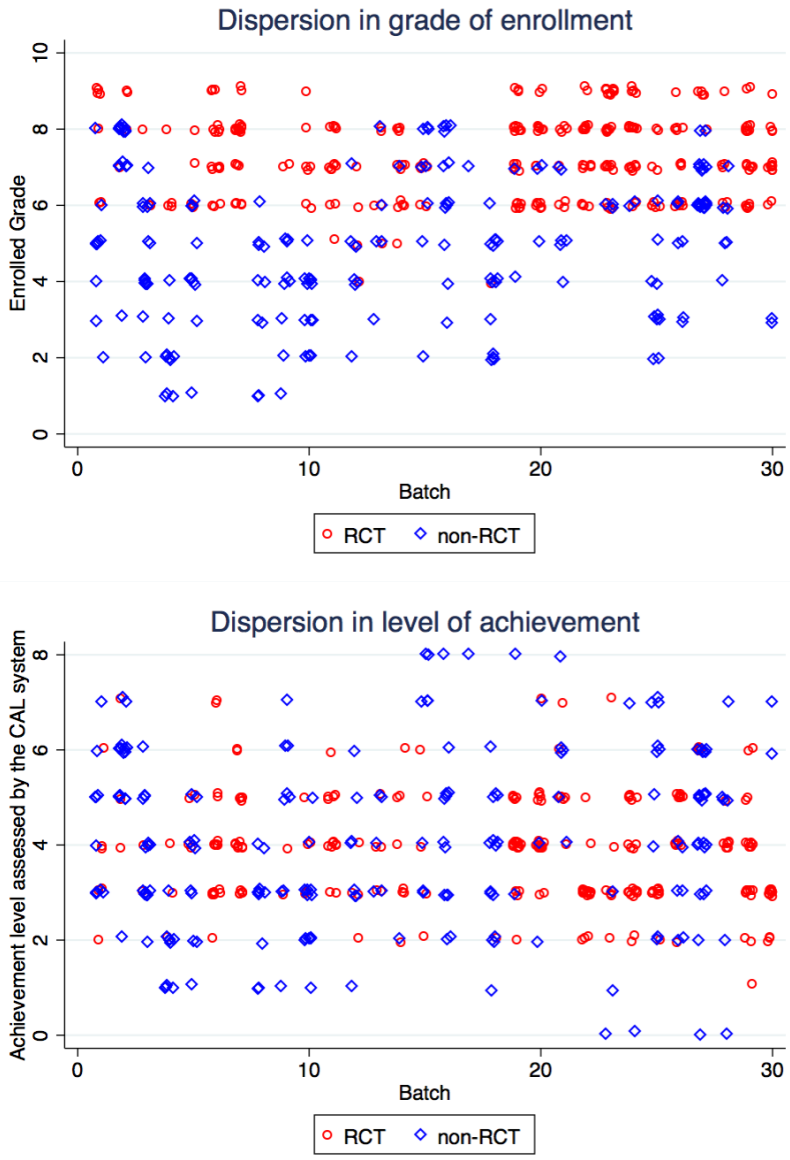
Note: The two panels above show mean test scores in Mathematics and Hindi respectively by each level of grade ability as assessed by the Mindspark CAL software at the beginning of the intervention (i.e. soon after the initial baseline) for students in the treatment group. Average test scores on our independently-administered assessments increase with CAL-assessed grade levels of achievement; this serves to validate that the two assessments capture similar variation and that Mindspark assessments of grade ability are meaningful. Only one student was assessed at Grade 1 level in math, and only 10 students at Grade 2 level in Hindi, the lowest categories in our sample in the two subjects. Consequently, scores are very noisy in these categories (and measurement error in the CAL assessments is also likely to be more severe).

Figure A.4: Distribution of questions administered by Mindspark CAL system



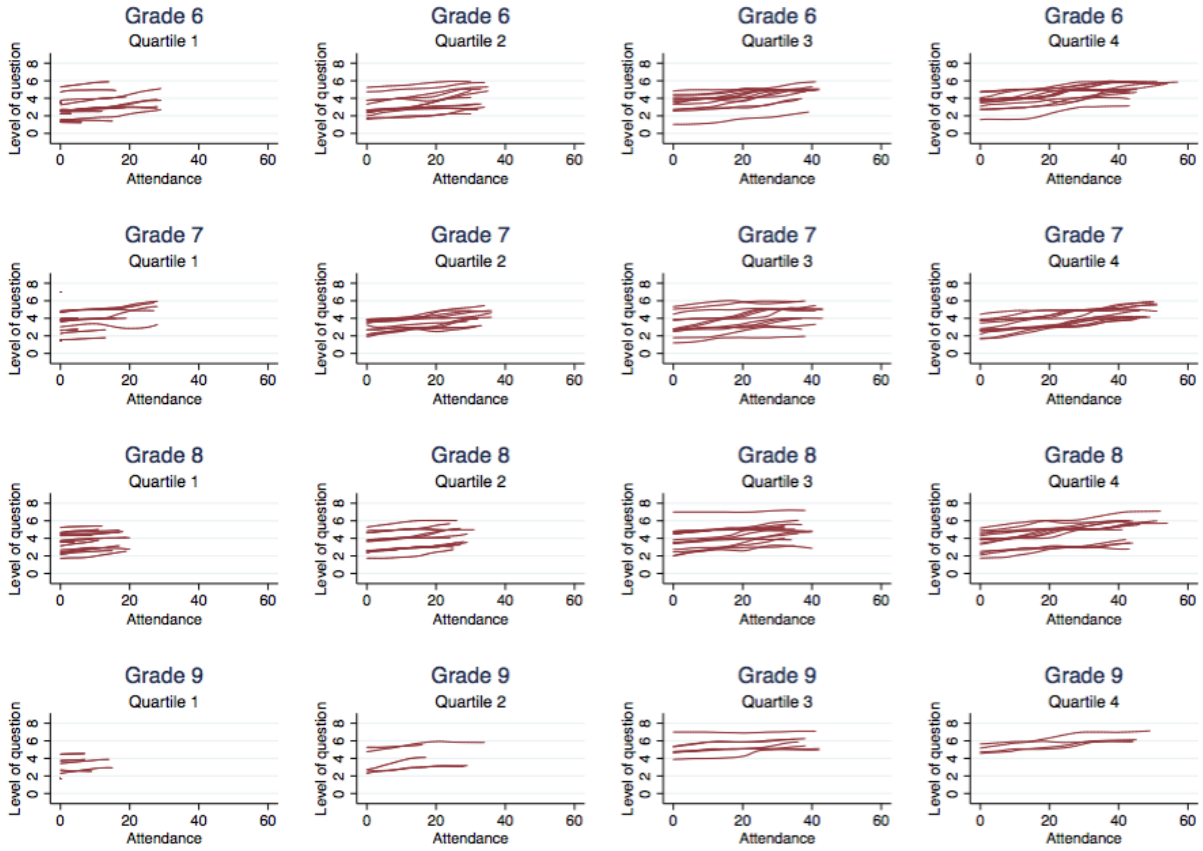
Note: The two panels above show the distribution, by grade-level, of the questions that were administered by the Mindspark CAL system over the duration of treatment in both math and Hindi. Note that in math, students received very few questions at the level of the grade they are enrolled in; this reflects 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 Figure 1).

Figure A.5: Composition of group instruction batches in Mindspark centers



Note: The two panels above show the composition of batches in Mindspark centers, by the grade students are enrolled in, and by their level of math achievement, as assessed by the Mindspark CAL system. We separately identify students in the treatment group from fee-paying students who were not part of the study but were part of the small group instruction in each batch. Note that, while our study is focused on students from grades 6-9, the centers cater to students from grades 1-8. Batches are chosen by students based on logistical convenience and hence there is substantial variation in grade levels and student achievement within each batch with little possibility of achievement-based tracking. This confirms that it would not have been possible to customize instruction in the instructor-led small group instruction component of the intervention.

Figure A.6: Learning trajectories of individual students in the treatment group



Note: Each line in the panels above is a local mean smoothed plot of the grade level of questions administered in Mathematics by the computer adaptive system against the days that the student utilized the Mindspark math software (Attendance). The panels are organized by the grade of enrolment and the within-grade quartile of attendance in Mindspark.

Table A.1: Comparing pre-program exam results of study participants and non-participants

	RCT	Non-study	Difference	SE	N(RCT)	N(non-study)
Math	0.13	-0.01	0.14	0.05	409	4067
Hindi	0.16	-0.02	0.17	0.05	409	4067
Science	0.09	-0.01	0.10	0.05	409	4067
Social Science	0.13	-0.01	0.15	0.05	409	4067
English	0.14	-0.01	0.15	0.05	409	4067

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

Table A.2: Intent-to-treat (ITT) effects with within-grade normalized test scores

	(1)	(2)	(3)	(4)
	Dep var: Standardized IRT scores (endline)			
	Math	Hindi	Math	Hindi
Treatment	0.38	0.23	0.38	0.23
	(0.068)	(0.066)	(0.069)	(0.071)
Baseline score	0.59	0.72	0.58	0.70
	(0.045)	(0.039)	(0.051)	(0.031)
Constant	0.33	0.20	0.33	0.19
	(0.047)	(0.046)	(0.034)	(0.035)
Strata fixed effects	Y	Y	N	N
Observations	523	525	523	525
R-squared	0.384	0.480	0.380	0.470

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

Table A.3: Heterogeneous effects on independent tests, by terciles of baseline achievement

	(1)	(2)	(3)	(4)
	Dep var: Proportion correct			
	Math		Hindi	
VARIABLES	At or above grade level	Below grade level	At or above grade level	Below grade level
Treatment	-0.030	0.059	0.095	0.10
	(0.054)	(0.020)	(0.043)	(0.026)
Treatment*Tercile 2	0.036	0.056	-0.053	-0.071
	(0.073)	(0.029)	(0.065)	(0.037)
Treatment*Tercile 3	0.13	0.023	-0.044	-0.079
	(0.080)	(0.032)	(0.062)	(0.033)
Tercile 1	0.24	0.45	0.39	0.49
	(0.045)	(0.017)	(0.041)	(0.022)
Tercile 2	0.26	0.46	0.38	0.58
	(0.037)	(0.015)	(0.030)	(0.018)
Tercile 3	0.39	0.54	0.55	0.67
	(0.042)	(0.018)	(0.037)	(0.019)
Baseline subject score	-0.015	0.069	0.087	0.084
	(0.032)	(0.010)	(0.023)	(0.011)
Observations	291	511	292	513
R-squared	0.096	0.371	0.301	0.433
	Total Treatment Effect by tercile (p-values in brackets)			
Tercile 1	-0.030	0.059	0.095	0.10
	[0.58]	[0.00]	[0.03]	[0.00]
Tercile 2	0.006	0.115	0.042	0.029
	[0.91]	[0.00]	[0.38]	[0.24]
Tercile 3	0.10	0.082	0.051	0.021
	[0.08]	[0.00]	[0.25]	[0.26]

Note: Robust standard errors in parentheses. Treatment is a dummy variable indicating a randomly-assigned offer of Mindspark voucher till March 2016. The total treatment effect by tercile is the sum of the coefficent on treatment and the interaction of the relevant tercile with the treatment. We report, in square brackets below the aggregate treatment effect in each tercile, p-values from an F-test of the hypothesis that this sum of the two coefficients is zero. The dependent variable and baseline scores are scaled as in Table 6

Table A.4: Correlates of attendance

VARIABLES	(1)	(2)	(3)	(4)
	Attendance (days)			
Female	3.90	2.65	3.03	4.06
	(3.90)	(3.92)	(3.88)	(3.88)
SES index	-3.33	-3.53	-3.47	-3.21
	(1.03)	(1.05)	(1.05)	(1.05)
Attends math tuition			-1.83	0.88
			(4.43)	(4.55)
Attends Hindi tuition			7.10	5.13
			(4.40)	(4.53)
Baseline math score		-0.99	-0.88	-0.81
		(2.17)	(2.24)	(2.24)
Baseline Hindi score		3.35	3.83	5.39
		(2.12)	(2.15)	(2.14)
Constant	46.6	47.5	45.3	43.7
	(3.40)	(3.42)	(3.79)	(3.78)
Grade Fixed Effects	N	N	N	Y
Observations	313	310	310	301
R-squared	0.038	0.046	0.056	0.120

Note: Robust standard errors in parentheses. This table shows correlates of days attended in the treatment group i.e. lottery-winners who had been offered a Mindspark voucher. Students from poorer backgrounds, and students with higher baseline achievement in Hindi, appear to have greater attendance but the implied magnitudes of these correlations are small. A standard deviation increase in the SES index is associated with a decline in attendance by about 3 days, and a standard deviation increase in Hindi baseline test scores is associated with an additional 5 days of attendance. We find no evidence of differential attendance by gender or by baseline math score.

Table A.5: Quadratic dose-response relationship

	(1)	(2)	(3)	(4)
	Full sample		Treatment group	
	Math	Hindi	Math	Hindi
Attendance (days)	0.0052	0.0079	0.0097	0.0070
	(0.0054)	(0.0053)	(0.0072)	(0.0073)
Attendance squared	0.000028	-0.000058	-0.000014	-0.000048
	(0.000073)	(0.000072)	(0.000083)	(0.000085)
Baseline subject score	0.58	0.71	0.62	0.68
	(0.042)	(0.040)	(0.061)	(0.052)
Constant	0.31	0.18	0.20	0.19
	(0.042)	(0.042)	(0.14)	(0.14)
Observations	535	537	264	265
R-squared	0.429	0.496	0.446	0.446

Note: Robust standard errors in parentheses. 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.6: Dose-response of subject-specific Mindspark attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Standardized IRT scores (endline)					
	IV estimates		OLS VA (full sample)		OLS VA (Treatment group)	
VARIABLES	Math	Hindi	Math	Hindi	Math	Hindi
Days of math instruction	0.018		0.019		0.022	
	(0.0029)		(0.0024)		(0.0047)	
Days of Hindi instruction		0.012		0.011		0.0084
		(0.0031)		(0.0026)		(0.0050)
Baseline score	0.56	0.68	0.58	0.71	0.61	0.68
	(0.038)	(0.036)	(0.041)	(0.039)	(0.060)	(0.052)
Constant			0.31	0.18	0.22	0.24
			(0.041)	(0.041)	(0.11)	(0.11)
Observations	535	537	535	537	264	265
R-squared	0.432	0.478	0.428	0.495	0.445	0.446
	19	19				
Angrist-Pischke F-statistic for weak instrument	1211	1093				
Diff-in-Sargan statistic for exogeneity (p-value)	0.12	0.80				
Extrapolated estimates of 45 days' treatment (SD)	0.81	0.54	0.855	0.495	0.99	0.378

Note: Robust standard errors in parentheses. Treatment group students who were randomly-selected for the Mindspark voucher 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 IV regressions which instrument attendance with the randomized allocation of a voucher and include fixed effects for randomization strata, Columns (3) and (4) present OLS value-added models for the full sample, and Columns (5) and (6) present OLS value-added models using only data on the lottery-winners. Scores are scaled here as in Table 2.

Table A.7: ITT estimates with inverse probability weighting

	(1)	(2)	(3)	(4)
	Dep var: Standardized IRT scores (endline)			
	Math	Hindi	Math	Hindi
Treatment	0.37	0.23	0.38	0.24
	(0.063)	(0.062)	(0.062)	(0.061)
Baseline score	0.59	0.71	0.57	0.68
	(0.041)	(0.040)	(0.038)	(0.037)
Constant	0.32	0.18	0.32	0.17
	(0.044)	(0.044)	(0.043)	(0.042)
Strata fixed effects	N	N	Y	Y
Observations	535	535	535	535
R-squared	0.405	0.487	0.454	0.535

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 voucher till March 2016. Results in this table are weighted by the inverse of the predicted probability of having scores in both math and Hindi in the endline; the probability is predicted using a probit model with baseline subject scores, sex of the child, SES index and dummies for individual Mindspark centers as predictors. Tests in both math and Hindi were designed to cover wide ranges of ability and to be linked across grades, as well as between baseline and endline assessments, using common items. Scores are scaled here as in Table 2.

Table A.8: Lee bounds estimates of ITT effects

	(1)	(2)
	Math	Hindi
Lower	0.309	0.183
	(0.092)	(0.102)
Upper	0.447	0.294
	(0.085)	(0.082)
Lower 95% CI	0.157	0.012
Upper 95% CI	0.587	0.43

Note: Analytic standard errors in parentheses. This table presents Lee(2009) bounds on the ITT effects of winning a voucher in both math and Hindi. We use residuals from a regression of endline test scores on baseline test scores (value-added) as the dependent variable, and scale scores as in Table 2, to keep our analysis of bounds analogous to the main ITT effects. The bounds are tightened using dummy variables for the Mindspark centres.

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

	(1)	(2)	(3)	(4)
	Math		Hindi	
VARIABLES	EI items	non-EI items	EI items	non-EI items
Treatment	0.11	0.075	0.055	0.044
	(0.013)	(0.011)	(0.017)	(0.011)
Baseline score	0.092	0.096	0.14	0.12
	(0.011)	(0.0084)	(0.0093)	(0.0052)
Constant	0.46	0.47	0.61	0.48
	(0.0064)	(0.0055)	(0.0082)	(0.0056)
Observations	537	537	539	539
R-squared	0.226	0.358	0.308	0.416

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

Table A.10: Treatment effect on take-up of other private tutoring

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

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

Appendix B Classroom Heterogeneity and Curricular Mismatch

As discussed in Sections 4.1 and 5.1, we conjecture that an important reason for the large effects we find is that the CAL software was able to accommodate the large heterogeneity in student learning levels within the same grade by personalizing instruction and teaching "at the right level" for all students. In this Appendix, we (a) provide evidence that the patterns in Figure 1 (a large fraction of students being behind grade-level standards and wide variation in academic preparation of students enrolled in the same grade) are present in other developing country settings as well, and (b) discuss qualitative evidence on pedagogical practice to show that the default instructional practice in these settings is to teach to the curriculum and textbook, which is likely to be above the learning levels of most students.

B.1 Comparing the distribution of achievement in our study sample with other samples

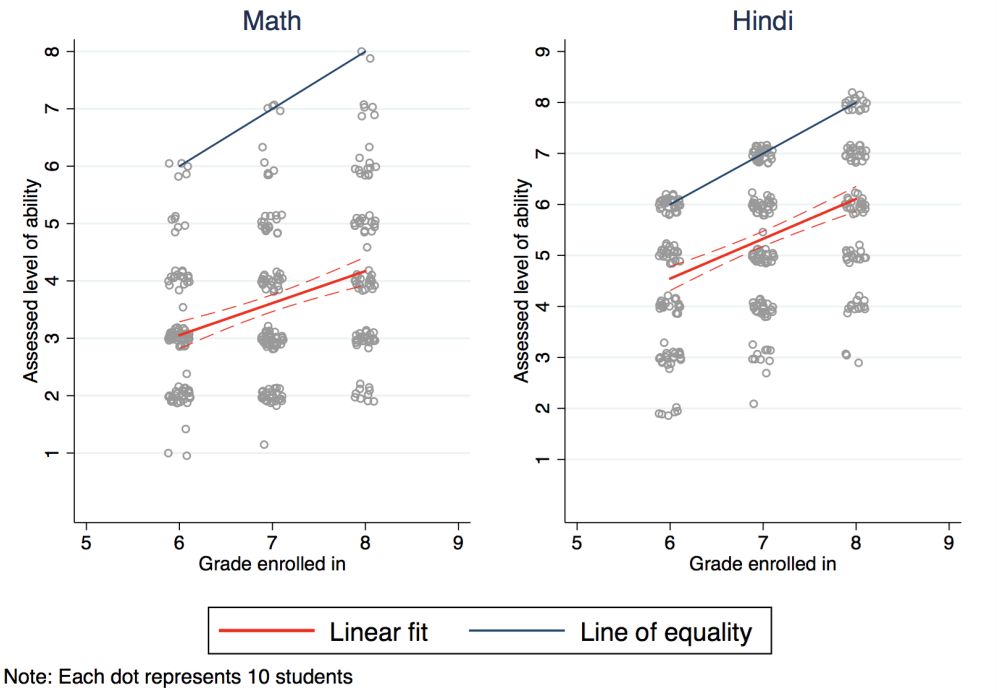
As mentioned in Section 4.1, an important advantage of the CAL data is the ability to characterize the mean and variance in grade-level preparation of students to produce the description shown in Figure 1. Yet, a limitation of the data in Figure 1 is that it comes from a self-selected sample of around 300 students in Delhi (though these students are quite similar to the other students in their school as seen in Figure A.1). We show now that these patterns are replicated in much larger and representative data sets of student learning in India.

B.1.1 Rajasthan

In September 2017, subsequent to our study, Educational Initiatives signed an agreement with the Government of the Indian state of Rajasthan to introduce the Mindspark software in 40 government schools in the state. This deployment was spread across urban and rural areas in 4 districts (Churu, Jhunjunun, Udaipur, and Dungarpur) spanning the northern and southern ends of the state of Rajasthan, and covered 3276 students across grades 6-8. A similar diagnostic exercise that informed Figure 1 was conducted for all these students and the data is presented in Figure B.1.

The patterns observed in Figure 1 are completely replicated in this larger and more representative (there was no student self-selection here) sample from a different state. Similar to the Delhi RCT sample, we see large absolute deficits against curricular standards (that grow in higher grades) and widespread dispersion within a grade. In math, the average Grade 6 student is 2.9 grade levels below curricular standards (compared to 2.5 grade levels below in Delhi), which rises to nearly 4 grade levels below by Grade 8 (similar to the sample in Delhi). In Hindi, the mean deficit in achievement compared to curricular standards is 1.5 grade levels

Figure B.1: Assessed achievement level vs. enrolled grade in 40 public schools in Rajasthan



Note: Each dot represents 10 students

in Grade 6, rising to 2 grade levels in Grade 8.[40] Thus, the patterns in the Rajasthan data are nearly identical to those in Delhi.

Since the Rajasthan data covers all students in the enrolled classes, we can also directly examine the within-classroom heterogeneity in learning levels (which we cannot see in Delhi because the sample there only includes students who signed up for the after-school Mindspark program). Using data from 116 unique middle-school classrooms across 40 schools, we see that the median classroom in these schools has a range of about 4 grade levels of achievement in both math and language. Consistent with the Delhi data, the dispersion is greater in higher grades and, at a maximum, we see a spread of up to 6 grade levels in achievement (Table B.1).

The Rajasthan data also allows us to decompose the within-grade variation in Figure B.1 into between and within classroom variation. Specifically, we find that classroom fixed effects account for 31% (19%) of the variation in grade-6 scores in math (Hindi), 24% (15%) of the variation in grade-7 scores in math (Hindi), and 19% (7%) of the variation in grade-8 scores in math (Hindi). Thus, the vast majority of the dispersion in learning levels in the same

[40] In 2017, Educational Initiatives modified the diagnostic test such that the maximum grade that a student would be assigned is the grade they are enrolled in. Thus, while students could advance to levels beyond curricular standards dynamically through the system, they could not start above grade level. This would understate the spread of achievement in the Rajasthan sample relative to the Delhi sample in Hindi (this is not an issue for math since almost no students are above grade level in math in Delhi).

Table B.1: Classroom-level heterogeneity in 40 schools in Rajasthan

		Mathematics		Hindi	
Grade		Range	p90 - p10	Range	p90 - p10
6	Mean	3.2	2.2	3.5	2.8
	Median	3	2	4	3
	Maximum	5	4	5	4
	N	40	40	40	40
7	Mean	4.1	3	3.9	3
	Median	4	3	4	3
	Maximum	6	5	5	4
	N	40	40	40	40
8	Mean	4.2	3	4.2	3.3
	Median	5	3	4.5	3.5
	Maximum	6	5	6	5
	N	36	36	36	36
Total	Mean	3.8	2.7	3.8	3
	Median	4	3	4	3
	Maximum	6	5	6	5
	N	116	116	116	116

grade seen in Figure Table B.1 is *within* classrooms and not between them, underscoring the challenge faced by teachers in effectively catering to such variation.

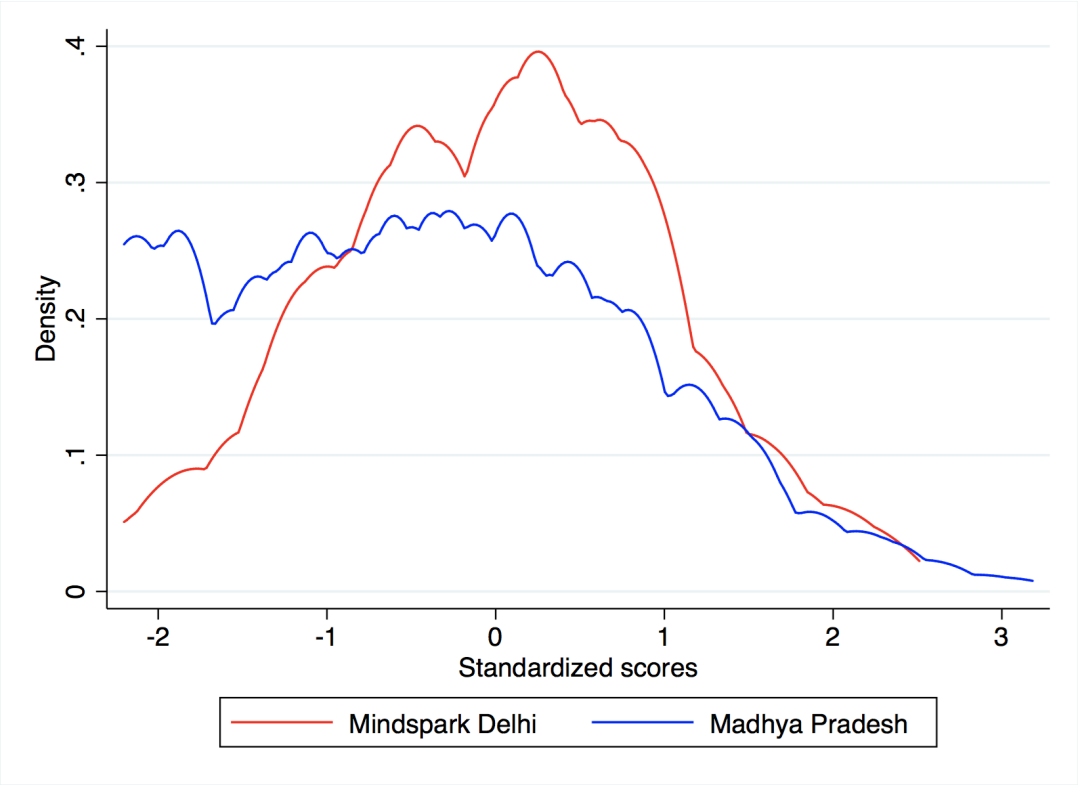
B.1.2 Madhya Pradesh

While data from the Mindspark CAL system from Rajasthan government schools provides the most direct comparison with the Delhi sample, an alternative comparison is possible using our independent assessments. In a separate contemporaneous study in the Indian state of Madhya Pradesh (MP) on the impact of a school-governance reform (Muralidharan and Singh, 2018), we administered a common subset of items from the Mindspark endline assessments. The MP sample consists of 2760 students in grades 6-8 (who were present on the day of the assessment) in a representative sample of government middle schools in 5 districts of Madhya Pradesh (MP).[41] Both the Delhi and the MP assessments were administered in February 2016. In Figure B.2 we present the distribution of achievement in the Madhya Pradesh sample with the control group in the Delhi Mindspark RCT on only the common items across the two studies; scores have been normalized to have a mean of zero and SD of one in the control group in the Delhi Mindspark RCT. The main results are that (a) the mean learning levels in MP are about 0.45 standard deviations below that in the Delhi sample, and (b) the standard deviation of the levels of student learning are about 25% greater than in the Delhi sample. Thus, both the key facts in Figure 1 (from the Delhi) sample of (a) low levels of learning, and

[41] Madhya Pradesh is the fifth-largest state in India by population with over 75 million inhabitants according to the 2011 Census. The state education system consists of over 112,000 schools.

(b) high variation within a grade are replicated in the MP data and appear to be even more pronounced.

Figure B.2: Distribution of achievement across the control group in the Mindspark RCT vs a representative sample of schools in Madhya Pradesh



B.1.3 Other countries and Indian states

There are two challenges in replicating the patterns of Figure 1 in other settings. First, most high-quality datasets on education in developing countries are from primary schools, whereas our focus in this paper is on post-primary grades. Second, while other datasets may allow the fact of variance in learning levels to be documented, the measures of learning are typically not linked to grade-level standards making it difficult to quantify the grade-level equivalent of learning gaps and variation (as we do above). We therefore focus on highlighting one key statistic on learning in developing countries, which is the fraction of students at the end of primary school (fifth or sixth grade) who are not able to read at the second grade level. The main advantage of this statistic is that is available in representative samples in several settings, and is also a meaningful measure of the phenomenon we are interested in – learning gaps (indicating a minimum of a three-year gap) and variation (since these students will be at least three years behind classmates who are at curricular standards). Table B.2 presents this number for several Indian states and other countries.

Table B.2: Learning standards in Grade 5 in India and selected other countries

State/Country	% Children who cannot read grade 2 level text	% Children who cannot do a division problem with a single-digit divisor	State/Country	% Children who cannot read grade 2 level text	% Children who cannot do a division problem with a single-digit divisor
Andhra Pradesh	55.1	37.2	Odisha	51.6	26.6
Arunachal Pradesh	25.5	19.0	Punjab	69.2	47.9
Assam	38.0	13.6	Rajasthan	54.2	28.2
Bihar	42.0	32.6	Tamil Nadu	45.2	21.4
Chattisgarh	55.9	23.0	Telangana	47.1	30.4
Gujarat	53.0	16.1	Tripura	51.0	19.9
Haryana	68.3	48.9	Uttar Pradesh	43.2	22.6
Himachal Pradesh	70.5	53.7	Uttarakhand	63.7	37.0
Jharkhand	36.4	23.5	West Bengal	50.2	29.0
Karnataka	42.1	19.7	**All India (rural)**	**47.8**	**25.9**
Kerala	69.2	38.6			
Madhya Pradesh	38.7	19.4	**Pakistan (rural)**	**52.1**	**48.4**
Maharashtra	62.5	20.3	Balochistan	41.7	39.9
Manipur	70.7	52.5	Punjab	65.0	59.6
Meghalaya	47.9	10.7	Sindh	36.6	24.3
Mizoram	46.0	27.7			
Nagaland	50.1	21.2	**Uganda**	**40.1**	**60.8**

Sources: Data for Indian states is taken from Pratham (2016), for Pakistan from SAFED (2017) and for Uganda from Uwezo (2016).

Note that students in Rajasthan perform slightly better than the national average for rural India, with several large states (such as Bihar, Madhya Pradesh and Uttar Pradesh) scoring substantially lower indicating that the challenges illustrated in Figure B.1 are likely to be even more severe in these settings. Similar patterns are also shown for two other countries (Pakistan, with major states shown separately, and Uganda) in which the grade of testing, the task tested, and the form of reporting is comparable with the ASER tests in India.

The pattern of large learning deficits, with significant heterogeneity within the same grade, is much more general. Table B.3 presents data from the World Development Report 2018 (World Bank, 2018) which consolidates data from 24 sub-Saharan countries, across three different assessments, to classify Grade 6 students by levels of competence in Reading and Mathematics. In most countries, a substantial proportion of students are classified as being "not competent" in mathematics.[42] However, there is substantial heterogeneity within the same grade in a country. In Kenya, for instance, about 30-40% of the sample is classified in

[42]For a concrete sense of what "not competent" means, in the PASEC assessment, this implies the inability to perform any but the most basic arithmetic operations with whole numbers (i.e. without demonstrating any knowledge of decimals or fractions or the ability to answer questions involving units of time, length or basic questions in geometry). In reading, it implies the inability to combine two pieces of explicit information in a text to draw simple inferences. In the SACMEQ assessments, "not competent" in reading implies the inability to link and interpret information located in various parts of the text; in math, it implies the inability to translate verbal or graphic information into simple word problems.

Table B.3: Classroom-level heterogeneity in 40 schools in Rajasthan

	Mathematics			Reading		
Country	Not competent	Low competence	High competence	Not competent	Low competence	High competence
All PASEC countries	57.6	24.7	17.7	61.6	25.1	13.3
All SACMEQ countries	36.8	18.4	44.8	63	20.2	16.8
Benin	48.3	29	22.7	60.2	29	10.8
Botswana	24.2	19.2	56.6	56.5	27.2	16.4
Burkina Faso	43.1	35.5	21.4	41.1	36.9	21.9
Burundi	43.5	49.1	7.4	13.2	46.8	39.9
Cameroon	51.2	24.7	24.1	64.6	23.7	11.8
Chad	84.3	12.8	3	80.9	16.1	3
Congo Rep.	59.3	23.5	17.1	71	23.1	5.9
Cote d'Ivoire	52	25.6	22.4	73.1	23.7	3.1
Kenya	19.8	19.6	60.6	38.3	32.1	29.6
Lesotho	52.5	25.5	22	81.1	13.6	5.3
Malawi	73.3	19.9	6.9	91.6	6.6	1.8
Mauritius	21.1	12.1	66.8	26.7	17.9	55.3
Mozambique	43.5	25	31.5	74.1	20.9	5
Namibia	38.7	25.5	35.8	81.7	12.2	6.1
Niger	91.5	6.4	2.1	92.4	6.3	1.4
Senegal	38.8	26.3	34.8	41.2	29.7	29.1
Seychelles	21.9	10.3	67.8	42.3	26.4	31.3
South Africa	48.3	14.7	37	69.2	15.4	15.5
Swaziland	7	20.7	72.2	44.3	37	18.7
Tanzania	10.1	12	77.9	43	25.5	31.5
Togo	61.6	22.6	15.8	52.5	27.9	19.7
Uganda	45.9	23.7	30.5	74.9	18	7.1
Zambia	72.6	14.9	12.4	91.8	6.5	1.7
Zimbabwe	37.2	20.7	42.1	57.2	22.6	20.2

Sources: This table draws upon figures presented in World Bank (2018), based on original data from SACMEQ (2007) , PASEC (2015) and the World Development Indicators.

each of the three bins of competence in mathematics (not competent, low competence, and high competence), highlighting the challenges of delivering a single program of instruction to all students in a classroom.

Taken together, the data presented in this section highlight that the two key patterns we highlight in Figure 1 of (a) large learning deficits relative to curricular standards, and (b) large heterogeneity in learning levels within the same grade, are typical of many developing country education systems.

B.2 Teaching to the curriculum

Inadequate and widely-dispersed academic preparation within a classroom would be a challenge for instruction in any setting. But it is made more severe if curricula and pedagogy are not responsive to this dispersion. Combined with the low general levels of achievement, this leads to substantial mismatch between the instruction delivered in the classroom and students' ability to engage with it. We see strong indirect evidence of this from our data.

First, we see that students scoring in the lowest-tercile of the within-grade baseline achievement distribution (who are at least a few grade levels behind the level of the curriculum) make no progress in absolute learning levels despite being enrolled in school – suggesting that the level of instruction within the classroom was too far ahead (and likely to have been at the level of the curriculum). Second, even though we see no program impact on average for treated students on the grade-level school tests, we see significant positive effects on these tests for students scoring in the top-tercile of the within-grade baseline achievement distribution. Since these students were exposed to Mindspark content that was closer to their grade level, it suggests that the school exams (and instruction in the school) are likely to have adhered to grade-level curricular standards.

In this section, we present additional qualitative evidence to show that classroom instruction in Indian schools closely tracks the textbook and curriculum, regardless of how far behind those standards most students may be. Two main sets of factors contribute to this.

B.2.1 Curriculum and syllabi

The first set relates to the prescribed curricula, syllabi and assessment. The way curricular standards are set and then transmitted in classroom teaching is largely determined by the (high-stakes) examination system, which serves later as a screening mechanism for future educational prospects and, eventually, white-collar jobs. In particular, it is not responsive to contextual factors about students' actual achievement or needs.[43] Although the National

[43]The National Focus Group on Curriculum, Syllabus and Textbooks, which underpinned the revised National Curriculum Framework in 2005, summarized the Indian education system as "largely a monolithic system perpetuating a kind of education which has resulted in a set of practices adopted for development

Curriculum Framework in 2005 did recommend unburdening the curriculum and making it more relevant, this has been hard to achieve in practice.[44] This focus on exam-oriented learning is particularly severe in middle and high schools, given major exam-based transition points after Grades 8 and 10. Given that post-primary education relies a great deal on foundational skills having been mastered, this focus means that a significant proportion of students are unable to engage with classroom instruction in a meaningful sense.[45]

B.2.2 The lack of differentiated instruction

The second set of issues relate to the ability and desire of teachers to address low and dispersed achievement in their classrooms of their own accord. While, in theory, it is possible for teachers to provide differentiated instruction to cater to widespread heterogeneity, there is no evidence that they do so. Sinha, Banerji and Wadhwa (2016) report, for instance, that 88% of primary and upper primary school teachers in Bihar believed that their main objective was to "complete the syllabus", even if nearly half of them agreed with or did not dispute the statement that "the textbooks are too difficult for children" (p. 24). Classroom observations at both primary and post-primary levels find consistently little evidence of differentiated or small-group instruction, with an overwhelming reliance on blackboard teaching and lecturing (Bhattacharjea, Wadhwa and Banerji, 2011; Sankar and Linden, 2014; World Bank, 2016; Sinha, Banerji and Wadhwa, 2016). Remedial instruction is also uncommon, and tracking of students into ability-based sections within school is made impractical in most public school settings in India because schools are small and rarely have more than one section per grade.[46]

In addition to reflecting the overall syllabus-determined orientation of the education system, the lack of remedial or differentiated instruction probably also reflects beliefs among some teachers about students' ability to learn. As Kumar, Dewan and K.Subramaniam (2012)

of curriculum, syllabus and textbooks that is guided by the patterns and requirements of the examination system, rather than by the needs determined by a mix of criteria based on the child's learning requirement, aims of education and the socio-economic and cultural contexts of learners." (NCERT, 2006)

[44] See e.g. Dewan and Chabra (2012) on the opposition to revising math curricula: "Even though the NCF is very clear on this issue, state functionaries continue to feel that reducing topics leads to loss of mathematical knowledge and children of their state are being deprived in this process. They also feel that such reductions will make their children unfit for various competitive examinations that they will take at the end of schooling."

[45] See e.g. Rampal and Subramaniam (2012) for a concrete example: "Mathematics at the upper primary level is premised on the ability to read and write numbers, and make sense of arithmetical expressions, as a starting point towards algebra. As children are not equipped to cope with this, classroom transaction gets reduced to children copying meaningless symbols from the blackboard, or from commercially available guidebooks in which the problems are worked out. Such classrooms where students cannot make sense of arithmetic expressions are not singular but fairly typical of classrooms catering to students from socioeconomically marginalised sections, or from rural backgrounds. They constitute a significant part of the student population."

[46] If anything, the opposite situation with the same teacher simultaneously teaching multiple grades is more typical. This is because the Indian government has prioritized universal access to school, resulting in several very small schools across rural India. The average enrollment in public schools in rural India is under 100 students across five primary grades, and the majority feature multi-grade teaching (Muralidharan et al., 2017).

discuss: "It is quite common for educators and administrators to believe that children from disadvantaged socio-economic backgrounds are incapable of learning mathematics, either because of an inherent lack of ability or because they do not have the cultural preparation and attitude to learning." Finally, it is not clear that, even had they wished to, teachers can effectively diagnose student errors and provide appropriate support. In a study of 150 secondary schools in two states (Madhya Pradesh and Tamil Nadu) in the 2014-2015 school year, it was found that language teachers were only able to identify student errors 50% of the time and math teachers were only able to do so 40% of the time (World Bank, 2016, p. 47). These challenges are not unique to India and similar findings of low teacher human capital and ability to support weaker students is also documented elsewhere; Bold et al. (2017), for example, use primary data from seven sub-Saharan African countries to document that "general pedagogical knowledge and the ability to assess students' learning and respond to that assessment is poor across the seven countries, with roughly only 1 in 10 teachers being classified as having minimum knowledge in general pedagogy and none having minimum knowledge in student assessment."

In sum, the core challenge of curriculum mismatch is general across the Indian education system. While direct evidence is scarce for other settings, it is likely that this challenge also generalizes to other developing country settings which are beset with low achievement and potentially over-ambitious curricula (see Pritchett and Beatty (2015)). Personalized instruction may also have significant potential for improving learning outcomes in these settings.[47]

[47]For experimental evidence, see Duflo, Dupas and Kremer (2011) which finds positive effects of tracking across the initial skill distribution and attributes it to the ability to customize instruction closer to skill levels of students within a classroom.

Appendix C Prior research on hardware and software

Tables C.1 and C.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).

C.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, Lin and Michko, 2003).

We implemented additional exclusions for each table. In Table C.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 C.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).

C.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. These decisions are non-trivial, as the specifications preferred by the authors of some studies (and reported in the abstracts) are only significant at the 10% level or only become significant at the 5% level after the inclusion of multiple covariates. Otherwise, we mentioned that a program had "no effect" on

the respective subject. Again, this decision is non-trivial because some of these studies were under-powered to detect small to moderate effects.

C.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 C.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 C.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 C.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	Cost of the voucher plus management costs 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, McNally and Silva (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 C.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, Onofa and Ponce (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*a*)	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*b*)	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, Markman and Rouse (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, Benson and Overman (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 D Mindspark software

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

D.1 Computer training

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

D.2 Diagnostic test

After the familiarization routine, students are presented with diagnostic tests in math and Hindi which are used by the Mindspark platform to algorithmically determine their initial achievement level (at which instruction will be targeted). Tests contain four to five questions per grade level in each subject. All students are shown questions from grade 1 up to their grade level. However, if students answer at least 75% of the questions for their corresponding grade level correctly, they can be shown questions up to two grade levels above their own.[48] If they answer 25% or less of the questions for one grade level above their actual grade, the diagnostic test shows no more questions. Initial achievement levels determined by the Mindspark system on the basis of these tests are only used to customize the first set of content that students are provided. Further customization is based on student performance on these content modules and does not depend on their performance on the initial diagnostic test (which is only used for initial calibration of each student's learning level).

D.3 Math and Hindi content

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

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

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

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

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

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

D.4 Personalization

D.4.1 Dynamic adaptation to levels of student achievement

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

Figure D.1 provides an illustration of how adaptability works. For example, a child could be assigned to the "decimal comparison test", an exercise in which he/she needs to compare two decimal numbers and indicate which one is greater. If he/she gets most questions in that test correctly, he/she is assigned to the "hidden numbers game", a slightly harder exercise in which he/she also needs to compare two decimal numbers, but needs to do so with as

little information as possible (i.e., so that children understand that the digit to the left of the decimal is the most important and those to the right of the decimal are in decreasing order of importance). However, if he/she gets most of the questions in the decimal comparison test incorrectly, he/she is assigned to a number of remedial activities seeking to reinforce fundamental concepts about decimals.

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

D.4.2 Error analysis

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

The idea that educators can learn as much (or perhaps more) from analyzing patterns of student errors than from their correct answers has a long tradition in education research (for instance, see Buswell and Judd (1925) and Radatz (1979) for discussions of the use of "error analysis" in mathematics education). Yet, implementing this idea in practice is highly non-trivial in a typical classroom setting for individual teachers. The power of 'big data' in improving the design and delivery of educational content is especially promising in the area of error analysis, as seen in the example below.

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

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

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

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

D.5 Feedback

The pedagogical approach favoured within the Mindspark system prioritizes active student engagement at all times. Learning is meant to build upon feedback to students on incorrect questions. Also, most questions are preceded by an example and interactive content that provide step-by-step instructions on how students should approach solving the question.

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

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

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

Figure D.1: Mindspark adaptability in math

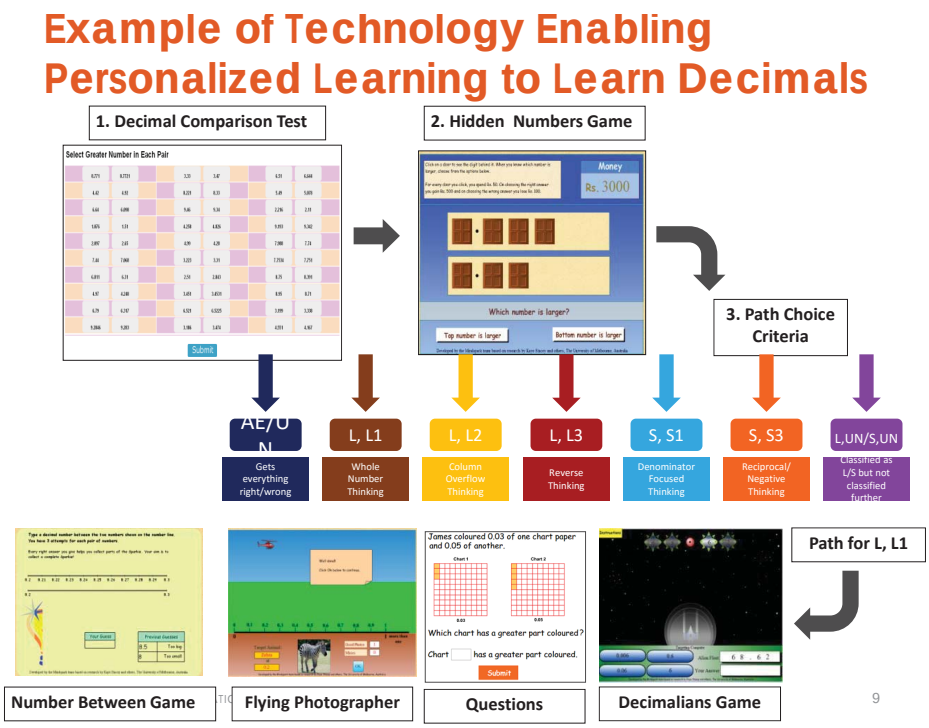
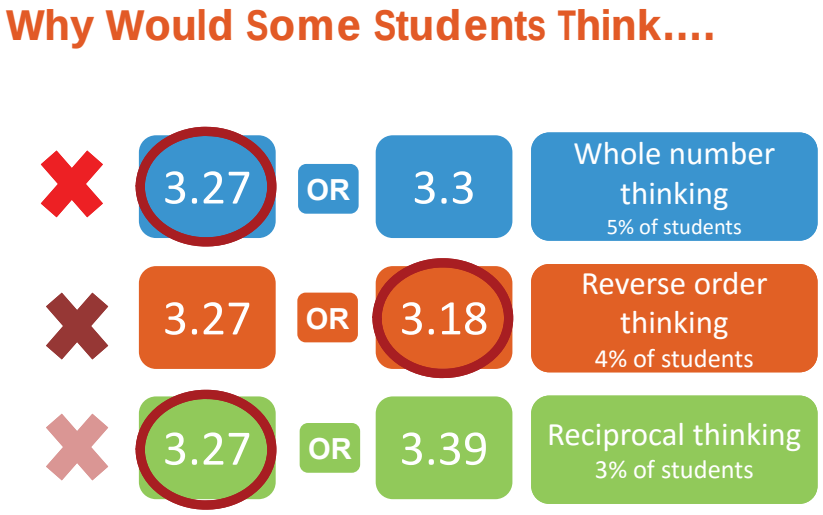


Figure D.2: Student errors in math



Appendix E Test design

E.1 Overview

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

E.2 Objectives of test design

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

Our secondary objective was to ensure that we measured 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.

E.3 Test content

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

In both subjects, we assembled the tests using publicly available items from a wide range of research assessments. In math, the tests drew upon items from the Trends in Mathematics and

Science Study (TIMSS) 4th and 8th grade assessments, OECD's Programme for International Student Assessment (PISA), the Young Lives student assessments administered in four countries including India, the Andhra Pradesh Randomized Studies in Education (APREStt), the India-based Student Learning Survey (SLS) and Quality Education Study (QES); these are collectively some of the most validated tests internationally and in 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.

E.4 Test booklets

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

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

E.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 (2013) and Das and Zajonc (2010).

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

Second, it allows us to robustly link all test scores on a common metric, even with only a partially-overlapping set of test questions, using a set of common items between any two assessments as "anchor" items. This is particularly advantageous when setting tests in samples with possibly large differences in mean achievement (but which have substantial common support in achievement) since it allows for customizing tests to the difficulty level of the particular sample but to still express each individual's test score on a single continuous metric. This is particularly important in our application in enabling us to compute business-as-usual value-added in the control group.[49]

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

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

$$P_g(\theta_i) = c_g + \frac{1 - c_g}{1 + exp(-1.7.a_g.(\theta_i - b_g))} \tag{3}$$

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).

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

E.6 Empirical distribution of test scores

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

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

E.7 Item fit

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

Figure E.3 plots the estimated Item Characteristic Curve (ICC) for each individual item in math and Hindi endline assessments along with the empirical fit for treatment and control groups separately. The fit of the items is generally quite good and there are no indications of differential item functioning (DIF) between the treatment and control groups. This indicates that estimated treatment effects do not reflect a (spurious) relationship induced by a differential performance of the measurement model in treatment and control groups.

Figure E.1: Distribution of raw percentage correct scores

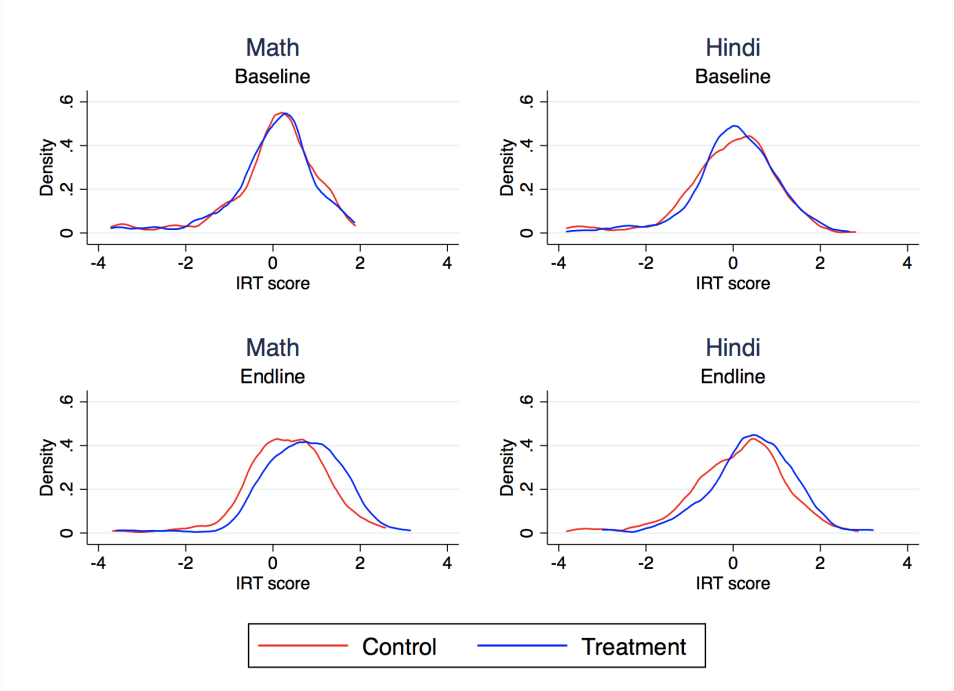


Figure E.2: Distribution of IRT scores, by round and treatment status

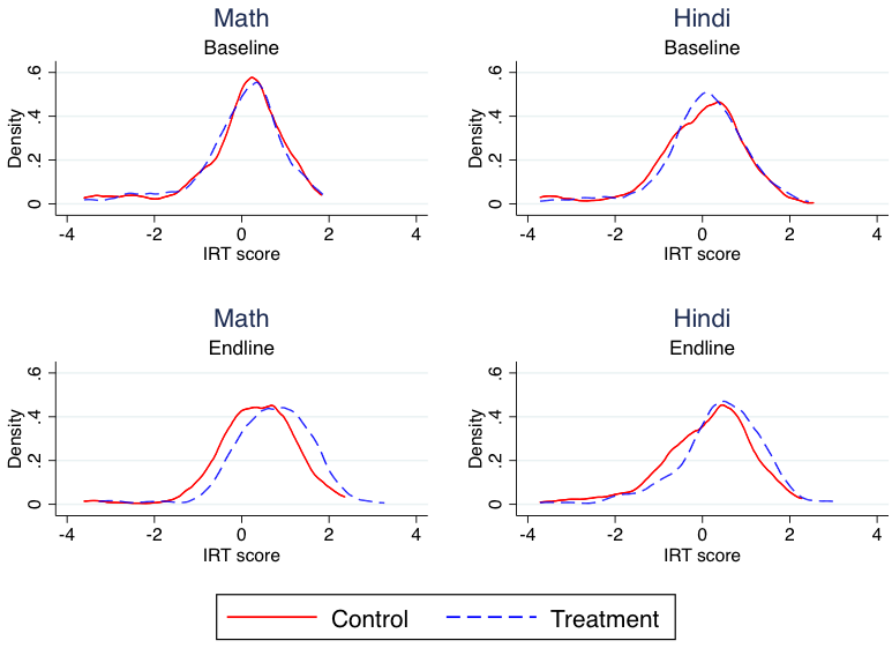
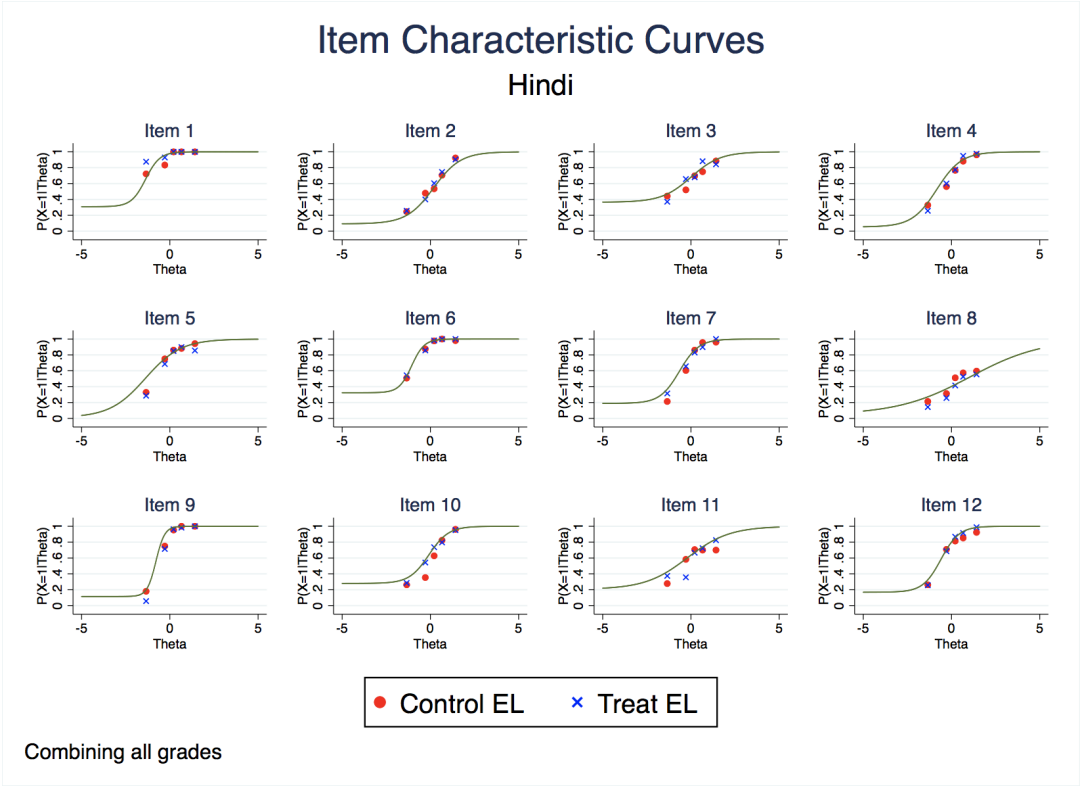
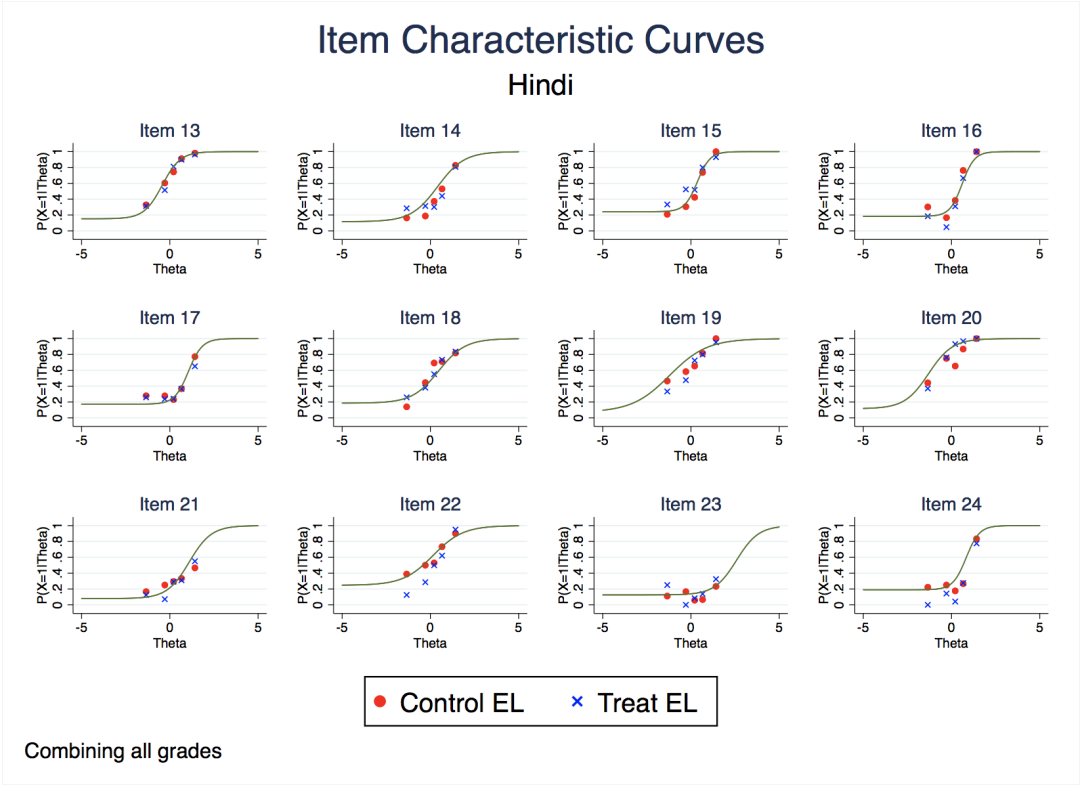


Figure E.3: Item Characteristic Curves: Hindi





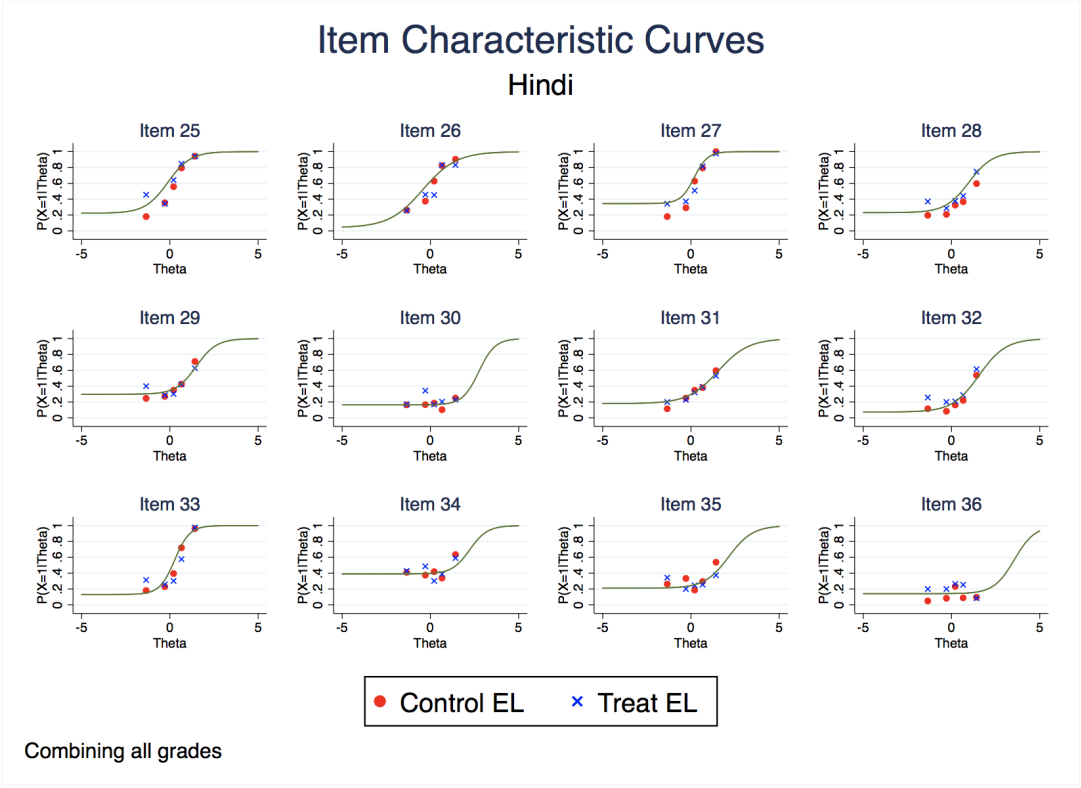
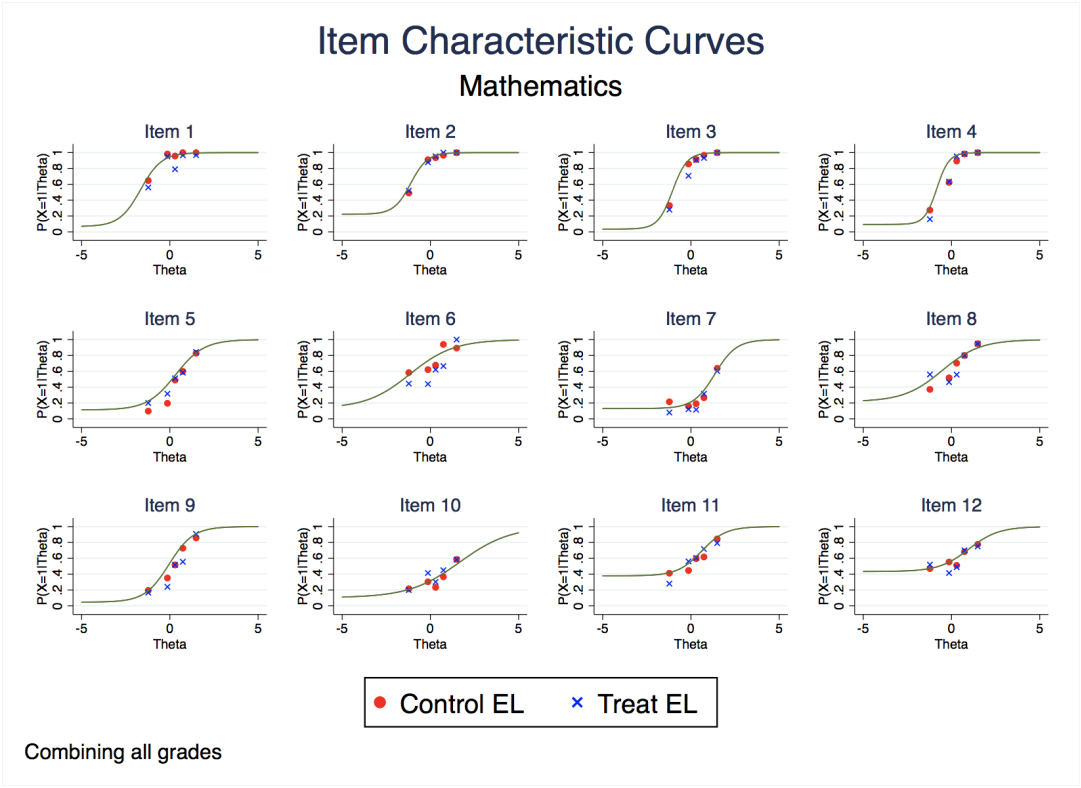
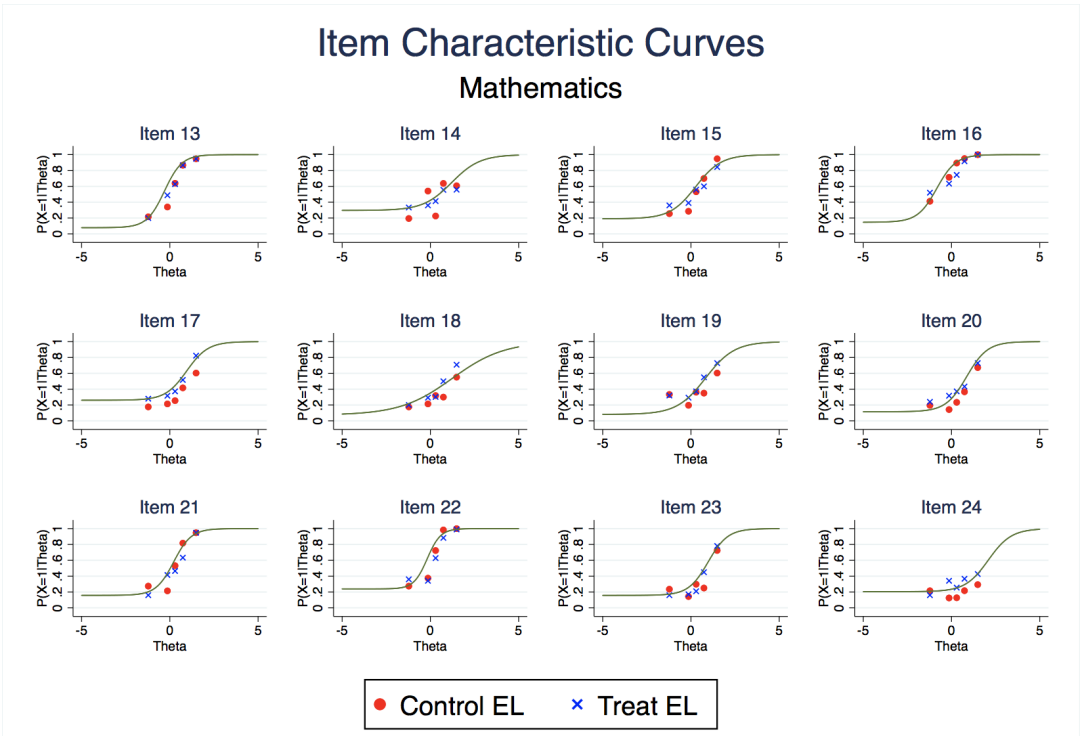
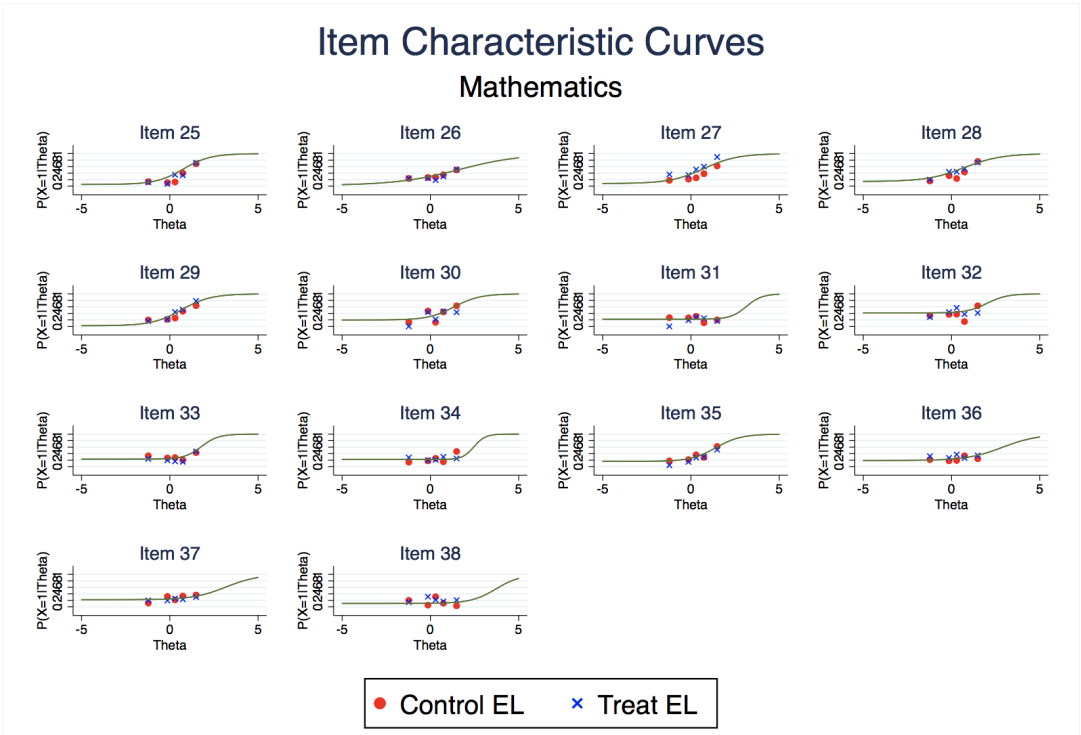


Figure E.4: Item Characteristic Curves: Math





Combining all grades



Combining all grades

Table E.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 E.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.