American Economic Review 2019, 109(4): 1426–1460
https://doi.org/10.1257/aer.20171112

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

By Karthik Muralidharan, Abhijeet Singh, and Alejandro J. Ganimian*

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 with 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* I21, I26, I28, J24, O15)

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 percent of students in grade 5 cannot read at the grade 2 level, despite primary school enrollment rates over 95 percent (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).

*Muralidharan: Department of Economics, University of California San Diego, 9500 Gilman Drive, La Jolla CA, NBER, and J-PAL (email: kamurali@ucsd.edu); Singh: Department of Economics, Stockholm School of Economics, Sveavagen 65, Stockholm, Sweden (email: abhijeet.singh@hhs.se); Ganimian: NYU Steinhardt School of Culture, Education, and Human Development, 246 Greene Street, New York, NY (email: alejandro.ganimian@nyu.edu). This paper was accepted to the *AER* under the guidance of Esther Duflo, Coeditor. We thank Esther Duflo, 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.

[†]Go to https://doi.org/10.1257/aer.20171112 to visit the article page for additional materials and author disclosure statement(s).

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 online 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 Level" 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 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 6 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

[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.

[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 IA and IVA for details).

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 one-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 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 reflecting an average attendance rate of 58 percent. Using the lottery as an instrumental variable for attendance (and additional assumptions discussed in Section IIID), we estimate that attending the Mindspark centers for 90 days (which corresponds to 80 percent attendance for one-half of 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 percent 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 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 5 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 centers could yield similar gains in one-tenth the time (one-half of 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 also targeted middle-school students in Delhi, 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 IVA).

Our results are directly relevant to policy debates on effective strategies to address the challenge of mismatch between student learning and the level of curriculum/pedagogy (which is a widespread problem in developing countries as

[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 IIIC).

documented in online 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 deliver the pedagogical advantages of tracking while mitigating several of its challenges (see discussion in Section IVC).

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 online 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.

[4] See, for example, Angrist and Lavy (2002), Barrera-Osorio and Linden (2009), Malamud and Pop-Eleches (2011), Cristia et al. (2012), and 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); Barrow, Markman, and Rouse (2009); and 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.

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 IVC), 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 I describes the intervention, and experimental design. Section II describes our data. Section III presents our main results. Section IV discusses mechanisms, costs, and policy implications. Section V concludes.

I. Intervention and Study Design

A. *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 one 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 online 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 online 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 nontrivial 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.

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

[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.

attend 6 days per week. The centers charged a (subsidized) fee of INR 200 (US$3) per month.[7]

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 IVA, 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.

B. *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

[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 percent of 11–17-year-olds attended paid private tutoring outside of school.

[8] These instructors were recruited based on two main criteria: (i) their potential to interact with children; and (ii) 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 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.

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).

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 one-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. 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.

Our study sample comprises the 619 students who completed the baseline tests and surveys. About 97.5 percent 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 (2014–2015) 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 nonparticipants (online Appendix Table A.1). However, there is near-complete common support in the pre-program test-score distribution of participants and nonparticipants (online Appendix 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 IIIC).

C. *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, socioeconomic status (SES), or baseline test scores (Table 1, panel A).[11] Among the 314 students offered a voucher for the program, the mean attendance rate was 58 percent (around 50 days out of a maximum possible of 86 days). The full distribution of attendance among lottery winners is presented in online Appendix Figure A.2,

[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 J-PAL 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 percent 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.

Table 1—Sample Descriptives and Balance on Observables

	Mean (treatment)	Mean (control)	Difference	Standard error	Observations (treatment)	Observations (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

Notes: Treatment and control groups refer to whether students were randomly assigned to receive an offer of a Mindspark voucher. Variables used in this table are from the baseline data collection in September 2015. The data collection consisted of two parts: (i) a self-administered student survey, from which demographic characteristics are taken and (ii) 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. *SES index* refers to a wealth index generated using the first component from a Principal Components Analysis consisting of indicators for ownership of various consumer durables and services in the household.

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 III.

Of the 619 students who participated in the baseline test, 539 (87 percent) also attended the endline test. The follow-up rate was 85 percent in the treatment group and 90 percent in the control group. This difference is significant at the 10 percent level and so we will present inverse probability-weighted estimates of treatment effects as well as Lee (2009) bounds of the treatment effect (Section IIIE). 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).

II. Data

A. *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 online Appendix E.

B. *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.

C. *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 2014–2015 school year (to compare the self-selected study sample with the full population of students in the same schools) and the 2015–2016 school year (to evaluate whether the treatment affected test scores on school exams).

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

D. *Student Data*

At the time of the baseline assessment, students answered a self-administered written student survey that collected basic information about their socioeconomic 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.

III. Results

A. *Learning Levels and Variation under the Status Quo*

Data from the Mindspark CAL system provide an assessment of the actual grade level of each student's learning 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 that 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 dispersion in achievement among students enrolled in the *same* grade: students in our sample span 5–6 grade levels in each grade.

In online 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 (i) the wide distribution of learning levels within a single grade is also seen in other settings and (ii) that a substantial proportion of students in grade 5 (toward the end of lower primary schooling in most countries) are often as much as 3 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

[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 online Appendix Figure A.3). Further details of the diagnostic test and benchmarking by the software are presented in online Appendix D.

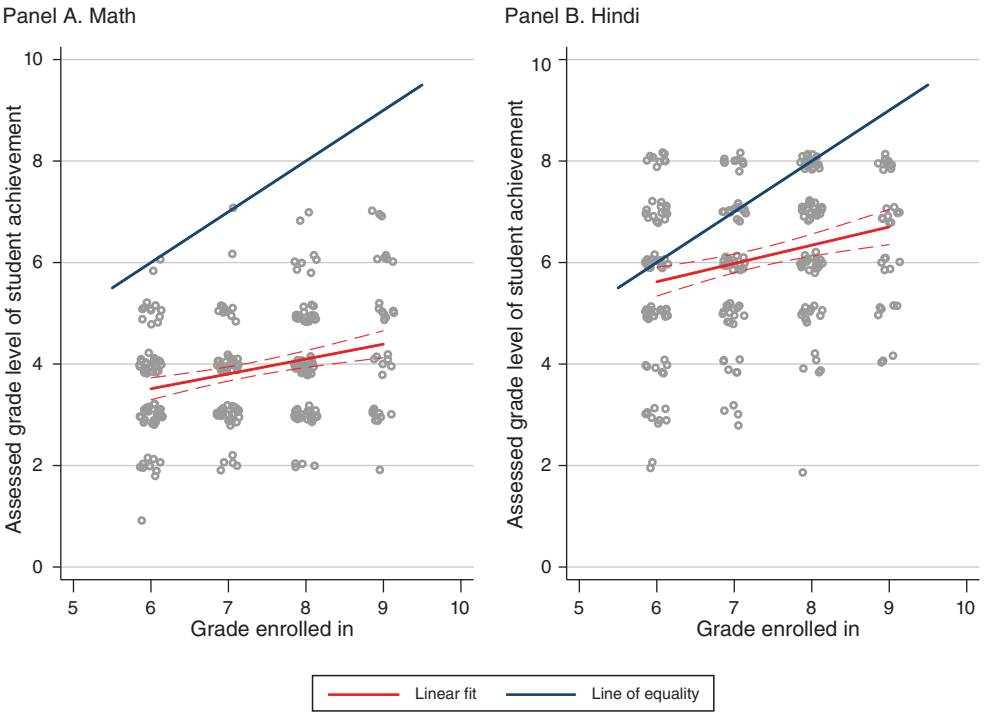

FIGURE 1. ASSESSED LEVELS OF STUDENT ACHIEVEMENT VERSUS CURRENT GRADE ENROLLED IN SCHOOL

Notes: 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: (i) there is a general deficit between average attainment and grade-expected norms; (ii) this deficit is larger in later grades; and (iii) within each grade, there is a wide dispersion of student achievement.

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.

B. *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

$$Y_{iks2} = \alpha_s + \gamma_s \cdot Y_{iks1} + \beta_s \cdot Treatment_i + \Phi_k + \epsilon_{iks2}, \quad (1)$$

where Y_{ikst} is student i's test score, in randomization stratum k, in subject s at period t (normalized to $\mu = 0$, $\sigma = 1$ on the baseline test); *Treatment* is an indicator variable

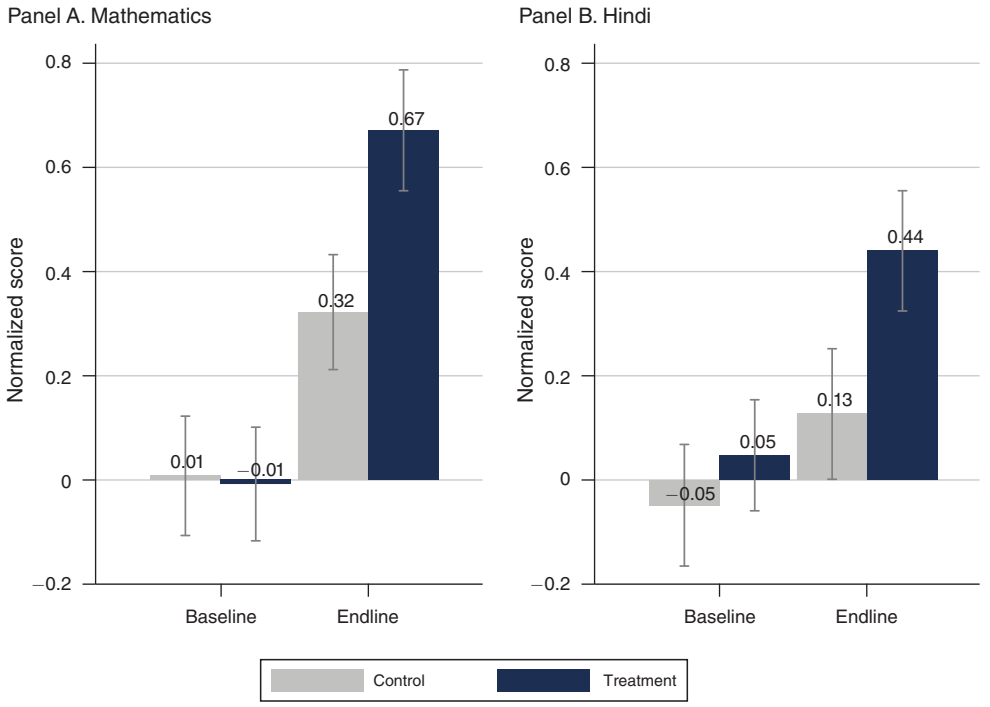

Figure 2. Mean Difference in Test Scores between Lottery Winners and Losers

Notes: 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 percent 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 0 and a standard deviation of 1 in the baseline. Whereas baseline test scores were balanced between lottery winners and lottery losers, endline scores are significantly higher for the treatment group.

for being a lottery winner; ϕ 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, columns 1–2). In columns 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. These are ITT results based on an average attendance of about 58 percent

[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 (online Appendix Table A.2).

Table 2—Intent-To-Treat (ITT) Effects in a Regression Framework

	Standardized IRT scores (endline)			
	Math (1)	Hindi (2)	Math (3)	Hindi (4)
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	Yes	Yes	No	No
Observations	535	537	535	537
R^2	0.403	0.493	0.397	0.473

Notes: Robust standard errors in parentheses. *Treatment* is a dummy variable indicating a randomly assigned offer of a Mindspark voucher. 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.

among lottery winners. We present IV results and estimates of a dose-response relationship in Section IIID.

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 percent increase on the easiest type of questions (arithmetic computation), determined by the proportion correctly answered in the control group, to a 38 percent increase on harder competencies such as geometry and measurement. Similarly, in Hindi, ITT effects range from a 6.4 percent gain on the easiest items (sentence completion) to a 17 percent gain on the hardest competence (answering questions based on interpreting and integrating ideas and information from a passage).

C. *Heterogeneity*

Heterogeneity by Student Characteristics.—We investigate whether ITT effects vary by gender, socioeconomic 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 nonparametric 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

Table 3—Treatment Effect by Specific Competence Assessed

	Proportion of questions answered correctly						
	Arithmetic computation	Word problems—computation	Data interpretation	Fractions and decimals	Geometry and measurement	Numbers	Pattern recognition
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Mathematics							
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^2	0.357	0.229	0.097	0.157	0.097	0.135	0.112

	Sentence completion	Retrieve explicitly stated information	Make straight-forward inferences	Interpret and integrate ideas and information
Panel B. Hindi				
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^2	0.182	0.380	0.309	0.136

Notes: Robust standard errors in parentheses. The tables 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 a Mindspark voucher. All regressions include randomization strata fixed effects.

Table 4—Heterogeneity in Treatment Effect by Gender, Socioeconomic Status, and Baseline Score

Covariates	Standardized IRT scores (endline)					
	Female		SES		Baseline score	
	Math	Hindi	Math	Hindi	Math	Hindi
	(1)	(2)	(3)	(4)	(5)	(6)
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^2	0.399	0.474	0.398	0.494	0.399	0.473

Notes: Robust standard errors in parentheses. *Treatment* is a dummy variable indicating a randomly assigned offer of a Mindspark voucher. 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.

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

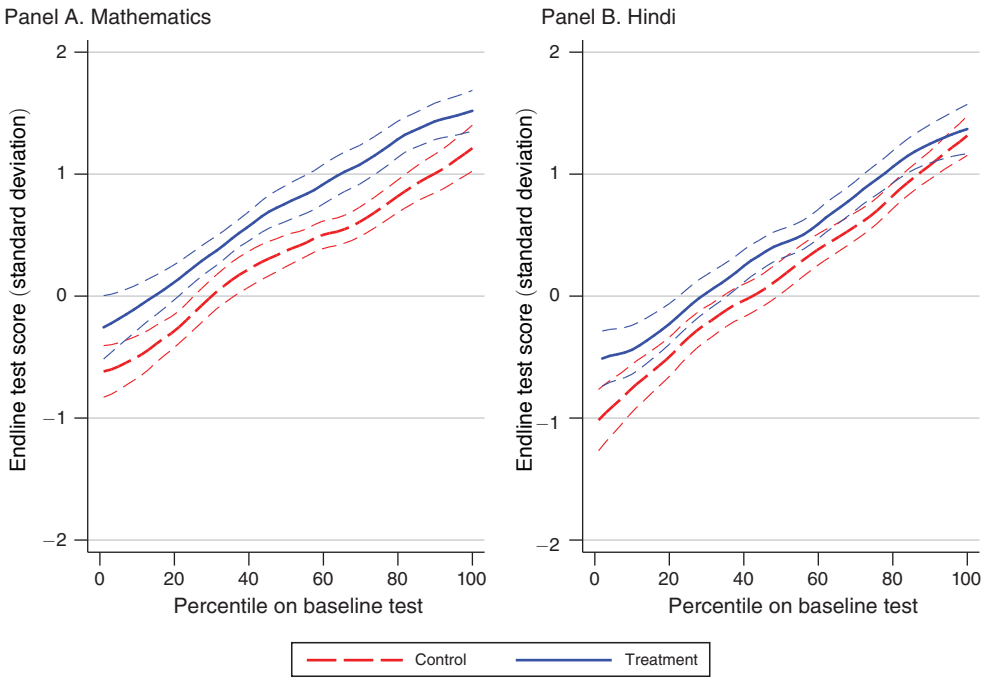

FIGURE 3. NONPARAMETRIC INVESTIGATION OF TREATMENT EFFECTS BY BASELINE PERCENTILES

Notes: 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 percent 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.

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 group over the same period (with coefficients close to 0 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 learned in the absence of

[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).

Table 5—Heterogeneity in Treatment Effect by Within-Grade Terciles

	Standardized IRT scores (endline)	
	Math	Hindi
	(1)	(2)
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^2	0.545	0.545

Notes: Robust standard errors in parentheses. *Treatment* is a dummy variable indicating a randomly assigned offer of a Mindspark voucher. Test scores are scaled as in Table 2.

treatment) were much larger for the weaker-performing students even though absolute gains are similar across all students (Figure 4).

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 online Appendix 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.

[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.

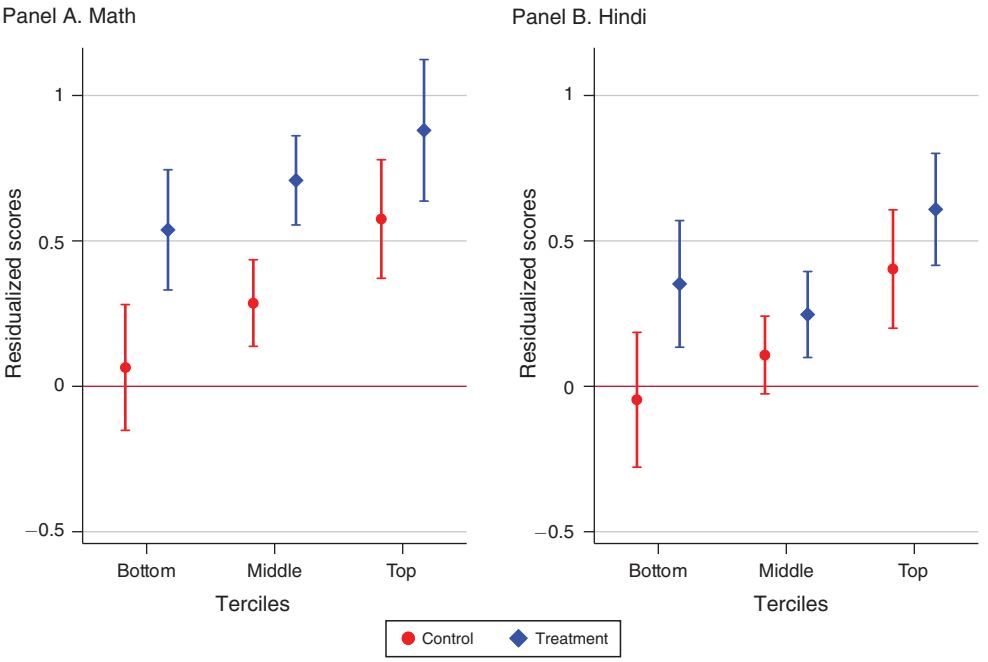

FIGURE 4. GROWTH IN ACHIEVEMENT IN TREATMENT AND CONTROL GROUPS

Notes: 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.

TABLE 6—TREATMENT EFFECT ON ITEMS LINKED TO GRADE LEVELS

	Proportion of questions answered correctly			
	Math		Hindi	
	At or above grade level (1)	Below grade level (2)	At or above grade level (3)	Below grade level (4)
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^2	0.029	0.346	0.250	0.399

Notes: 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

	Standardized test scores					
	Hindi	Math	Science	Social sciences	English	Aggregate
	(1)	(2)	(3)	(4)	(5)	(6)
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^2	0.190	0.073	0.121	0.177	0.144	0.210

Notes: 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 a Mindspark voucher. 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.

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.

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 from 0.2–0.5σ), with gains on 4 out of 5 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 (online Appendix 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

[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 $\mu = 0$, $\sigma = 1$ in each grade/subject in the control group.

[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 practicing past exam questions) that could lead to higher performance on school exams in the short term.

Table 8—Heterogeneous Effects on School Tests, by Terciles of Baseline Achievement

	School test scores					
	Hindi (1)	Math (2)	Science (3)	Soc. Sc. (4)	English (5)	Aggregate (6)
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^2	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]

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

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 would 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 online Appendix E for further details on

[21] On our tests, gains in Hindi are larger (and only statistically significant) for the bottom tercile (online Appendix 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.

test design for our study, and Muralidharan (2017) for a detailed discussion on test construction for education research in general.

D. *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 percent among lottery winners.[22] In this section, we present LATE estimates of the impact of 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

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

where Y_{ist} is defined as previously, *Attendance* is the number of days a student logged into 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, columns 1 and 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, p. 435). Using these IV estimates to predict the effect of varying the number of days attended requires further assumptions about (i) 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 (ii) 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 equation (3) and the ordinary least squares (OLS) estimates using a value-added (VA) specification (Table 9, columns 3 and 4), which suggests that the average treatment effects and local average treatment effects (ATE and LATE) may be similar here. For both math and Hindi, the *p*-value from the difference-in-Sargan test (similar to a

[22] About 13 percent of lottery winners attended for 1 day or less. The mean attendance among the rest was 57 days (around 66 percent). Online Appendix Figure A.2 plots the distribution of attendance among lottery winners, and online Appendix Table A.4 presents correlations of attendance among lottery winners with various baseline characteristics.

Table 9—Dose-Response of Mindspark Attendance

	Standardized IRT scores (endline)					
	IV estimates		OLS VA (full sample)		OLS VA (treatment group)	
	Math (1)	Hindi (2)	Math (3)	Hindi (4)	Math (5)	Hindi (6)
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^2	0.431	0.479	0.429	0.495	0.446	0.445
Angrist-Pischke F-statistic for weak instrument	1,207	1,244				
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

Notes: 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 zero percent 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, 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.

Hausman test, but allowing for heteroskedasticity) testing equivalence of OLS and IV results is substantially greater than 0.1 (columns 1 and 2).[23]

Finally, the constant term in the OLS VA specifications (corresponding to 0 attendance) is similar when estimated using the full sample and when estimated using only the data in the treatment group (Table 9, columns 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 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, columns 3–6), adding a quadratic term does not improve fit, and the quadratic term is not significant (see online Appendix Table A.5). A linear dose-response

[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. 2017).

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

[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.

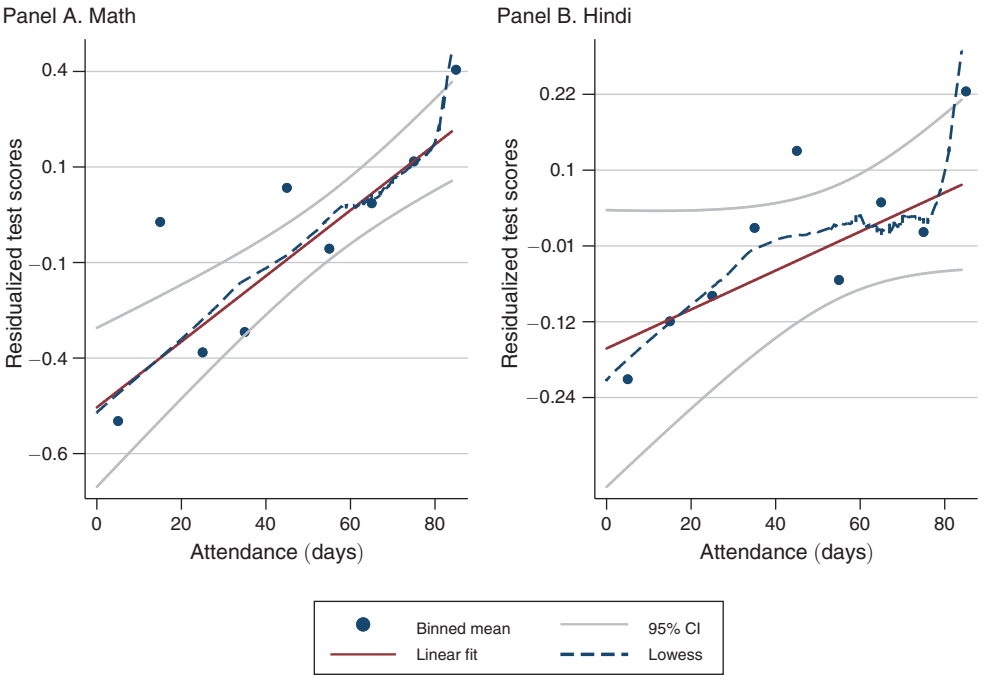

FIGURE 5. DOSE RESPONSE RELATIONSHIP

Notes: 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 nonparametric plot.

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 one-half of a school year with 80 percent 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 online

[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.

Appendix 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 online Appendix Table A.6).

E. *Robustness*

Attrition.—Since the difference in attrition between the treatment and control groups is significant at the 10 percent 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 (online Appendix 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 (online Appendix Table A.8).

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 (online Appendix Table A.9).

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 online Appendix 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.

IV. Discussion

A. *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 6 days of instruction per week, for 3 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 nonprofit 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 and 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 that the CAL software presented students with material that is either *at their grade level or at adjacent*

[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 the 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.

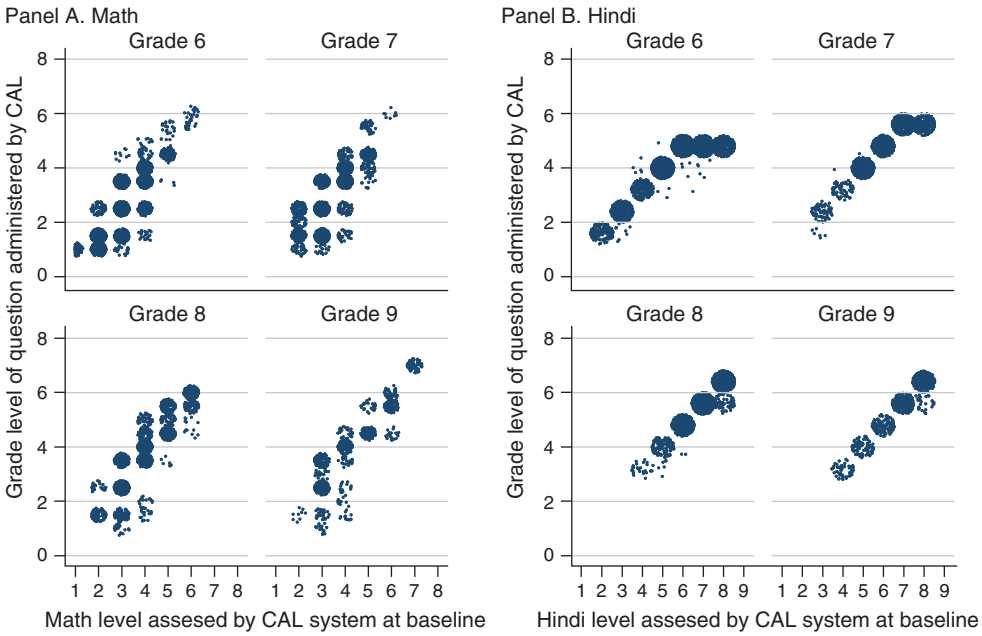

FIGURE 6. PRECISE CUSTOMIZATION OF INSTRUCTION BY THE MINDSPARK CAL PROGRAM

Notes: 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.

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 nonparametric 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 online Appendix 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

[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 IA). 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.

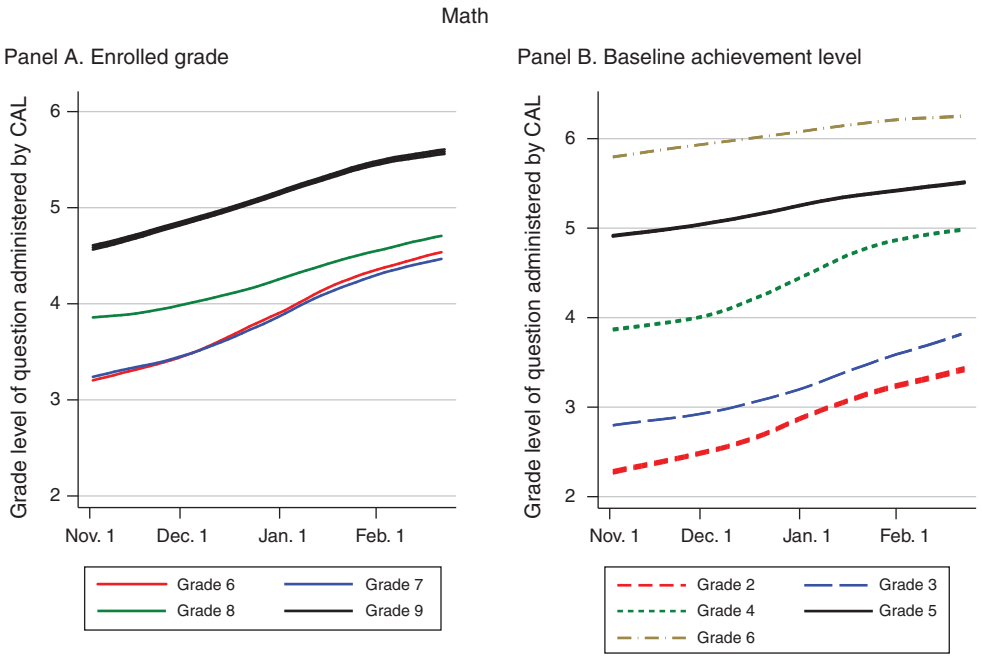

FIGURE 7. DYNAMIC UPDATING AND INDIVIDUALIZATION OF CONTENT IN MINDSPARK

Notes: 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. Panel A presents separate lines by the actual grade of enrollment. Panel B presents separate lines by the level of achievement assessed at baseline by the CAL software. Note that 95 percent 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.

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 suggest 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 online Appendix 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 setup that would make very little sense for group instruction.[32]

Finally, note that the patterns of test-score results we present in Section IIIC 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

[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.

software taught) and not on grade-level questions (which were not taught by the CAL software). This is also consistent with the 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.

B. *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) suggests 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 (US$22) in 2014–2015; 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 overestimates 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 (US$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 percent 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 US$4 per month (including hardware costs). Above a scale of 1,000 schools, the per-pupil marginal costs (software maintenance and

technical support) are about US$2 annually, which is a small fraction of the US$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 5 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 of a year would yield similar gains (in one-tenth of 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 percent 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.

C. *Policy Implications*

Despite the large test-score gains we find, parental demand for Mindspark centers was low in the absence of (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 percent capacity (and thereby demonstrate parental willingness to pay at least the subsidized price). In practice, enrollment levels were considerably below this target, and the centers had to shut down because philanthropic funding for the subsidies ended.[35] Thus, 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, policymakers (especially in developing countries) have mainly concentrated on providing computer hardware without

[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 five years will lead to ten times greater test-score gains, but simply note that the gains observed over five years in Muralidharan (2012) were achieved in one-tenth of 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.

commensurate attention to using technology to improve pedagogy.[36] Our results (combined with the review of evidence in online 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 labeled 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.

V. 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 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 address other constraints to 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

[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 textbooks for the corresponding grade.

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. These "big data" on student achievement also offer much potential of their 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 instructional time; and whether it is delivered without supervision, under the supervision of current teachers, or under the supervision of third parties (e.g., 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 nonroutine 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

[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.

[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.

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 noncognitive 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 nontrivial 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 ways of delivering these benefits at a larger scale.

REFERENCES

Andrabi, Tahir, Jishnu Das, Asim Ijaz 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, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871–919.

Angrist, Joshua, and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–42.

Angrist, Joshua, and Victor Lavy. 2002. "New Evidence on Classroom Computers and Pupil Learning." *Economic Journal* 112 (482): 735–65.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279–333.

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." NBER Working Paper 22746.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

Banerjee, Abhijit, and Esther Duflo. 2012. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: 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." Unpublished.

Barrera-Osorio, Felipe, and Leigh Linden. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia." World Bank Policy Research Working Paper 4836.

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, James, and Priya Mukherjee. 2016. "Pricing of Private Education in Urban India: Demand, Use and Impact." Unpublished.

Bertanha, Marinho, and Guido W. Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." NBER Working Paper 20773.

Betts, Julian R. 2011. "The Economics of Tracking in Education." In *Handbook of the Economics of Education*, Vol. 3, edited by Eric Hanushek, Stephen Machin, and Ludger Woessmann, 341–81. Amsterdam: 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.

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

Bulman, George, and Robert W. Fairlie. 2016. "Technology and Education: Computers, Software, and the Internet." In *Handbook of the Economics of Education*, Vol. 5, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 239–80. Amsterdam: Elsevier.

Carrillo, Paul, Mercedes Onofa, and Juan Ponce. 2010. "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." Inter-American Development Bank Working Paper IDB–WP–223.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125 (1): 1–45.

Cristia, Julian, Pablo Ibarraran, Santiago Cueto, Ana Santiago, and Eugenio Severin. 2012. "Technology and Child Development: Evidence from the One Laptop Per Child Program." Inter-American Development Bank Working Paper IDB–WP–304.

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

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132 (4) 1593–640.

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.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–74.

Fujiwara, Thomas. 2015. "Voting Technology, Political Responsiveness, and Infant Health: Evidence from Brazil." *Econometrica* 83 (2): 423–64.

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education*, Vol. 5, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 653–744. Amsterdam: Elsevier.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–64.

Hirshleifer, Sarojini R. 2015. "Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance." Unpublished.

Jack, William, and Tavneet Suri. 2014. "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution." *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): 915–43.

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.

Lai, Fang, Renfu Luo, Linxiu Zhang, Xinzhe Huang, and Scott Rozelle. 2015. "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing." *Economics of Education Review* 47: 34–48.

Lai, 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* 5 (2): 208–31.

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." Stanford Rural Education Action Program Working Paper 237.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–102.

Linden, Leigh. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." Unpublished.

Malamud, Ofer, and Cristian Pop-Eleches. 2011. "Home Computer Use and the Development of Human Capital." *Quarterly Journal of Economics* 126 (2): 987–1027.

Mo, Di, Linxiu Zhang, Renfu Luo, Qinghe Qu, Weiming Huang, Jiafu Wang, Yajie Qiao, Matthew Boswell, and Scott Rozelle. 2014. "Integrating Computer-Assisted Learning into a Regular Curriculum: Evidence from a Randomised Experiment in Rural Schools in Shaanxi." *Journal of Development Effectiveness* 6 (3): 300–323.

Muralidharan, Karthik. 2012. "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India." Unpublished.

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

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106 (10): 2895–929.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India: Dataset." *American Economic Review.* https://doi.org/10.1257/aer.20171112.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120 (546): F187–203.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130 (3): 1011–66.

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

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: 276–88.

Radatz, Hendrik. 1979. "Error Analysis in Mathematics Education." *Journal for Research in Mathematics Education* 10 (3): 163–72.

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): 323–38.

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.

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