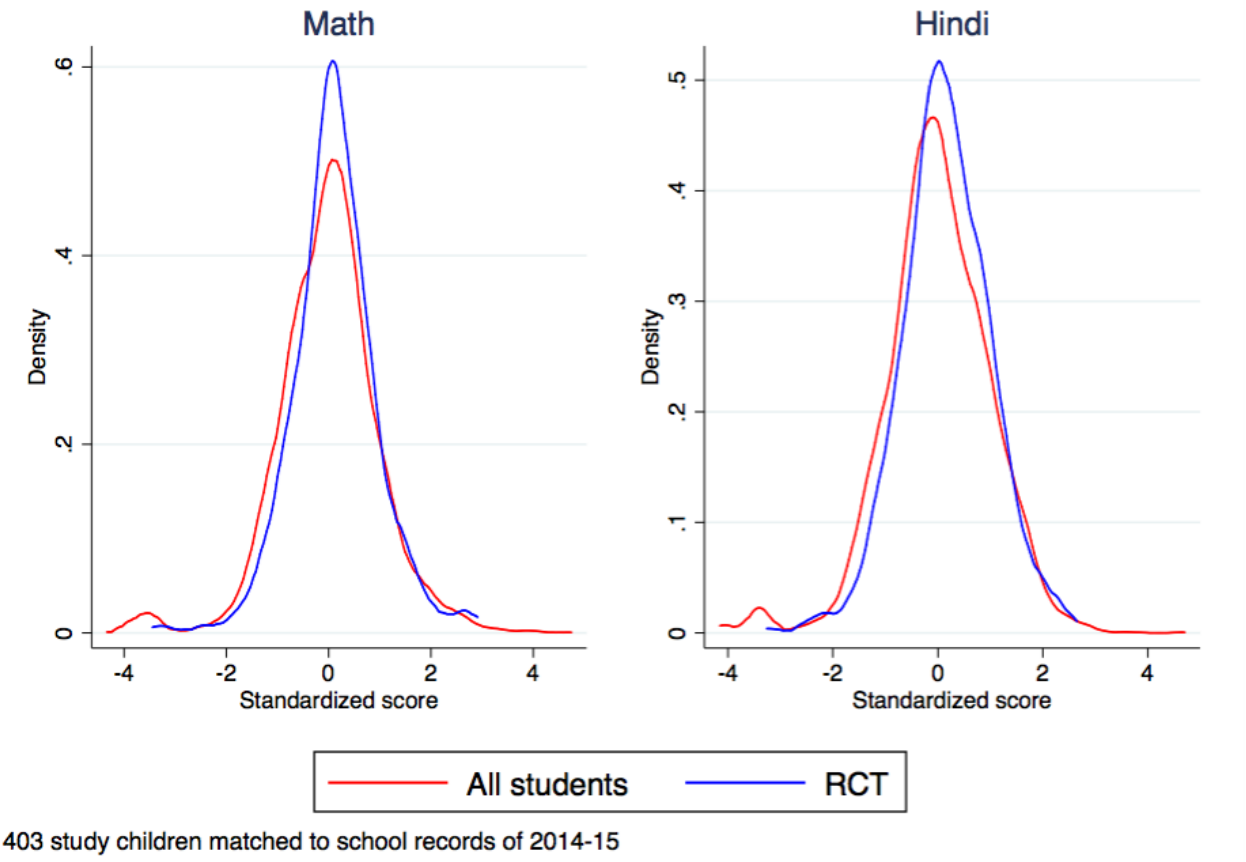
Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India

Karthik Muralidharan, Abhijeet Singh, Alejandro Ganimian

Appendices (for Online Publication Only)

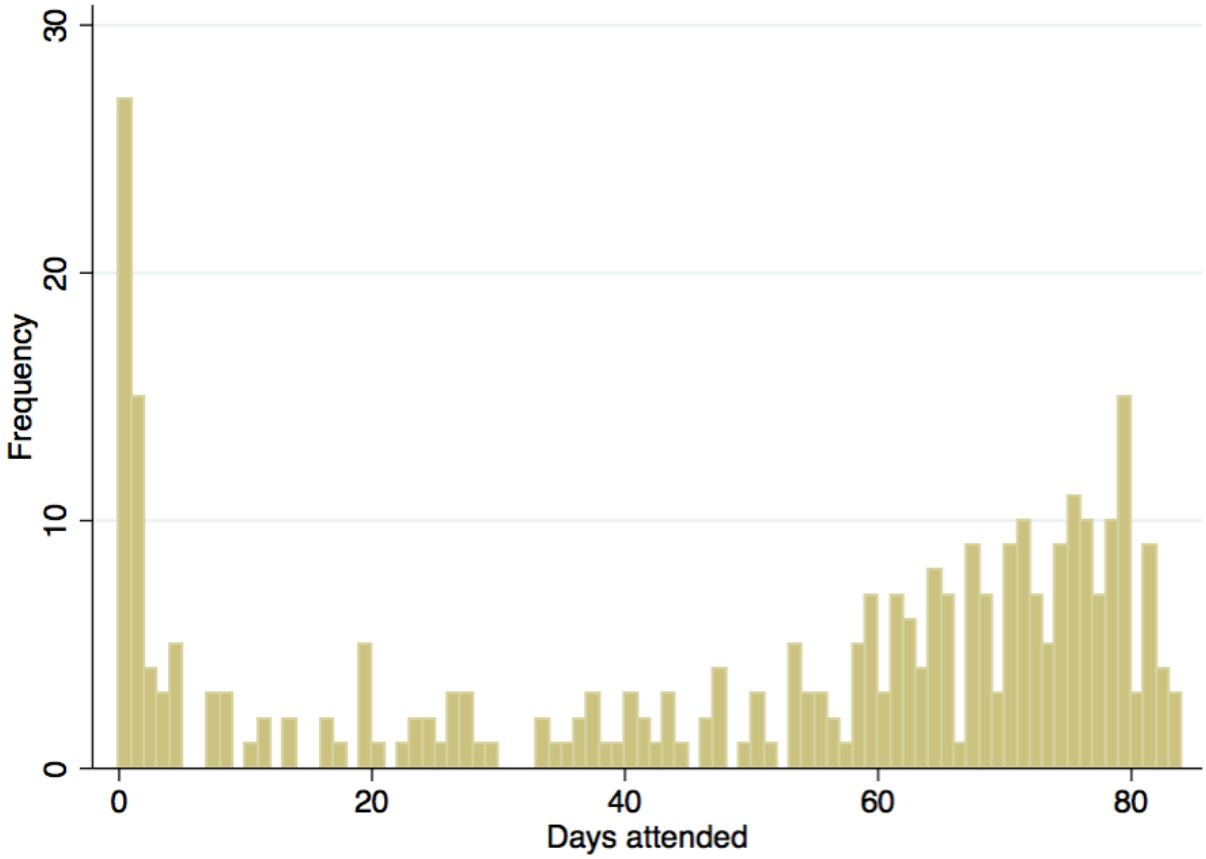
Appendix A Additional figures and tables

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



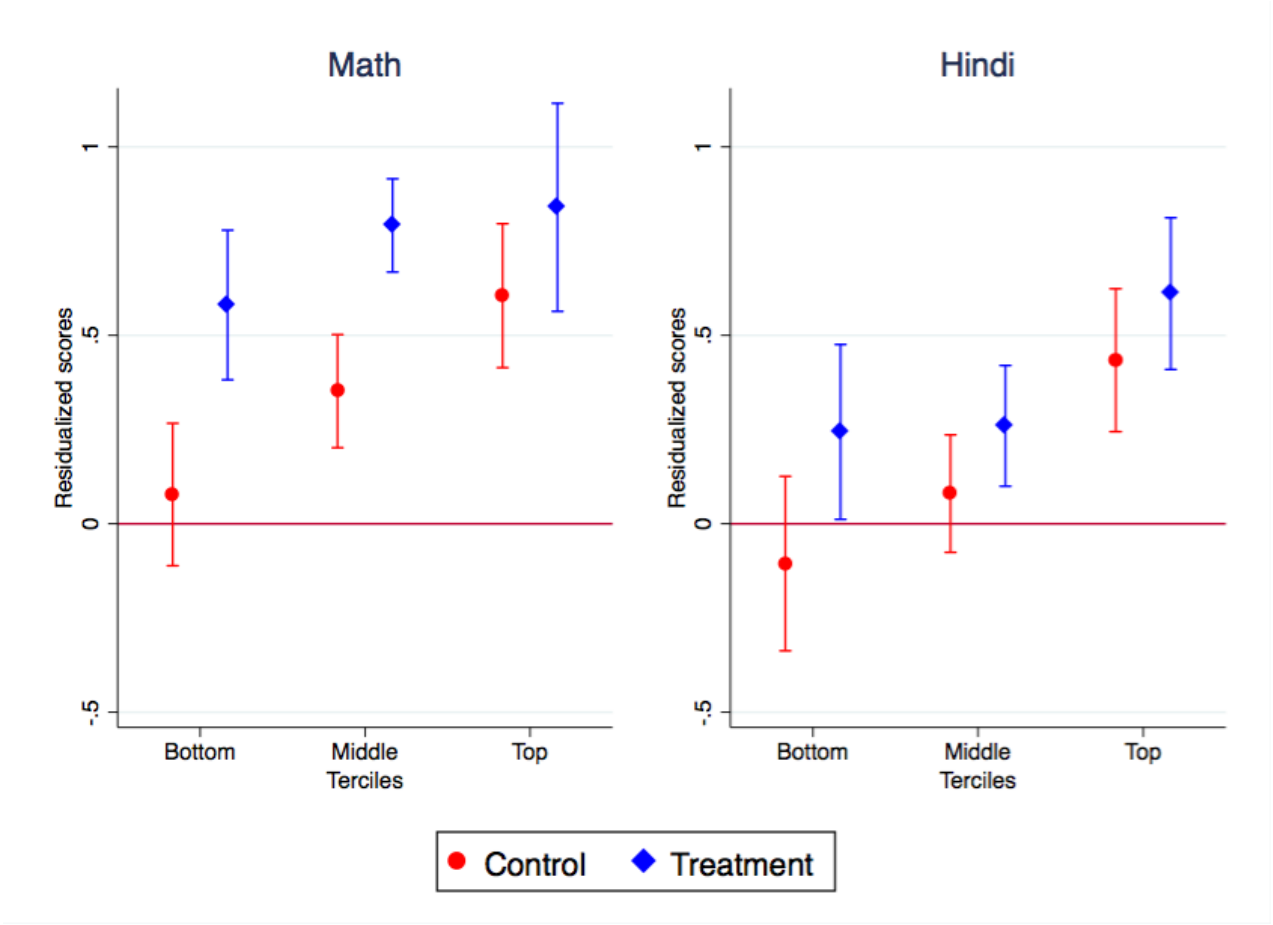
Note: The panels compare the final scores for the 2014-15 school year, i.e. the pre-program academic year, for study participants and non-participants. Test scores have been standardized within school*grade cells. The study participants are positively selected into the RCT in comparison to their peers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

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



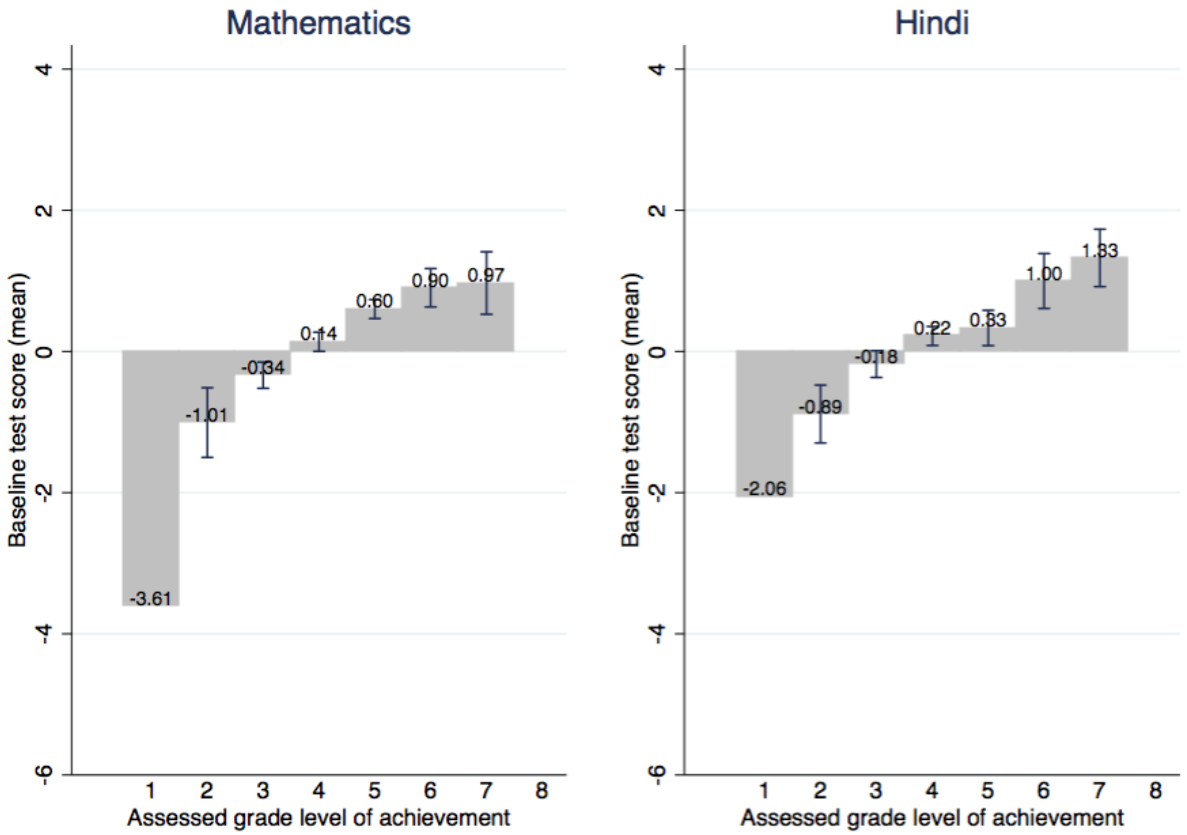
Note: This figure shows the distribution of attendance in the Mindspark centers among the lottery-winners. Over the study period, the Mindspark centers were open for 86 working days.

Figure A.3: Growth in achievement in treatment and control groups



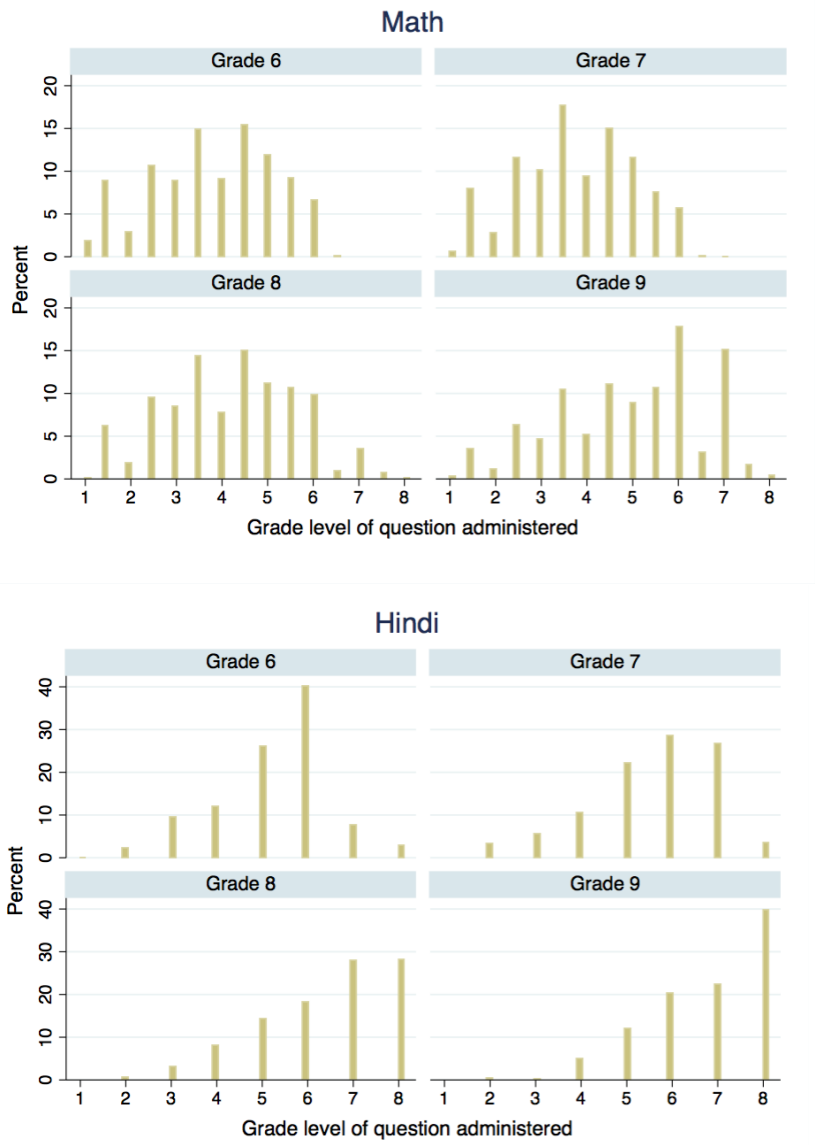
Note: This figure shows the growth in student achievement in the treatment and control groups in math and Hindi, as in Table 5. Students in the treatment group see positive value-added in all terciles whereas we cannot reject teh null of no academic progress for students in the bottom tercile in the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

Figure A.4: Comparison of Mindspark initial assessment of grade-level of student achievement with (independent) baseline test scores



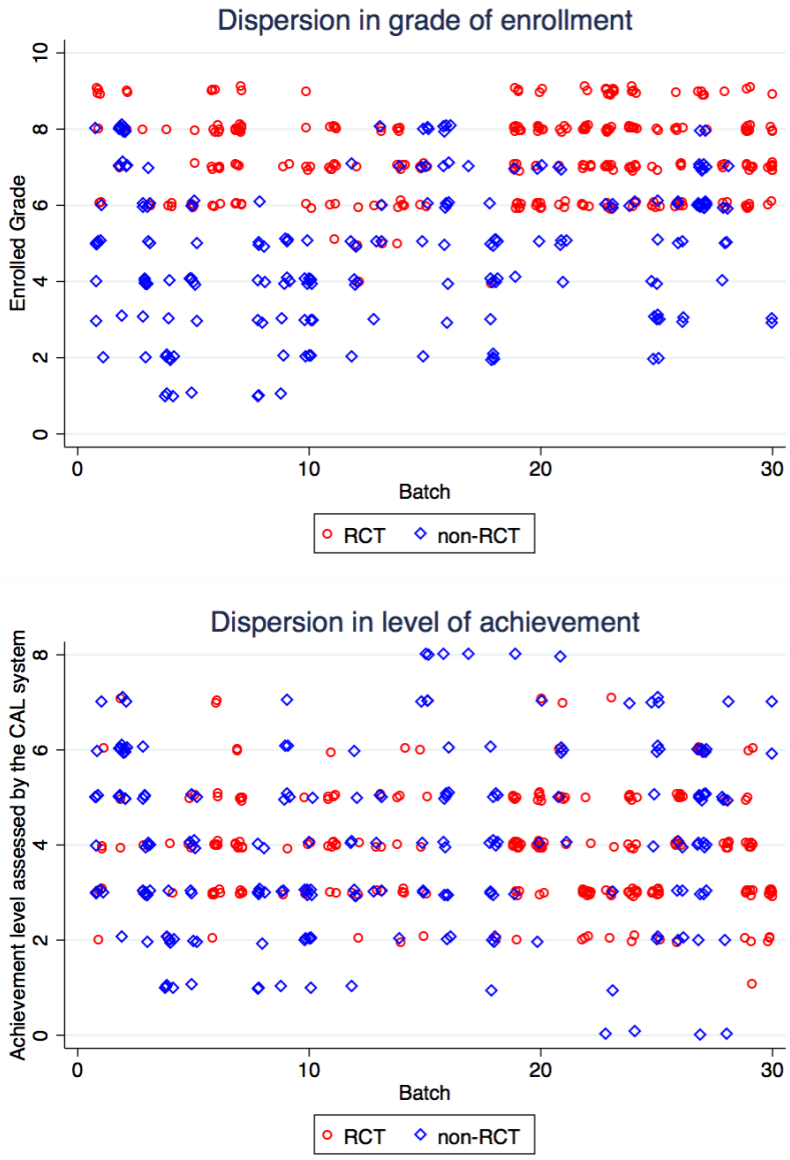
Note: The two panels above show mean test scores in Mathematics and Hindi respectively by each level of grade ability as assessed by the Mindspark CAL software at the beginning of the intervention (i.e. soon after the initial baseline) for students in the treatment group. Average test scores on our independently-administered assessments increase monotonically with each level of grade ability; this serves to validate that the two assessments capture similar variation and that the Mindspark assessments of grade ability are meaningful.

Figure A.5: Distribution of questions administered by Mindspark CAL system



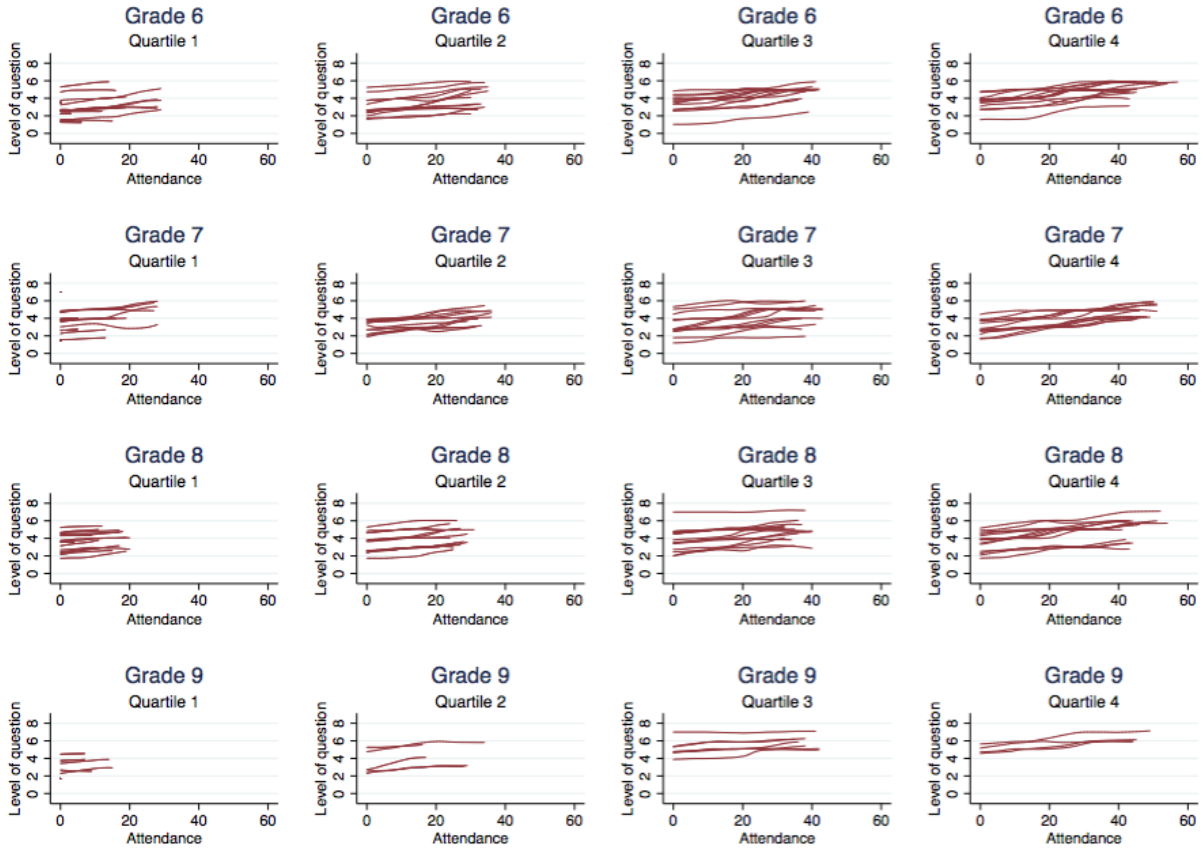
Note: The two panels above show the distribution, by grade-level, of the questions that were administered by the Mindspark CAL system over the duration of treatment in both math and Hindi. Note that in math, students received very few questions at the level of the grade they are enrolled in; this reflects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

Figure A.6: Composition of group instruction batches in Mindspark centers



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

Figure A.7: Learning trajectories of individual students in the treatment group



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

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

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

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

Table A.2: ITT estimates with within-grade normalized test scores

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

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

Table A.3: Correlates of attendance

VARIABLES	(1)	(2)	(3)	(4)
	Attendance (days)			
Female	3.81	2.51	2.89	4.00
	(3.90)	(3.93)	(3.89)	(3.90)
SES index	-3.26***	-3.49***	-3.43***	-3.19***
	(1.04)	(1.07)	(1.06)	(1.06)
Attends math tuition			-1.95	0.62
			(4.41)	(4.53)
Attends Hindi tuition			7.27*	5.32
			(4.38)	(4.50)
Baseline math score		-1.07	-0.99	-0.59
		(2.05)	(2.11)	(2.09)
Baseline Hindi score		3.66*	4.17**	5.49***
		(2.06)	(2.10)	(2.10)
Constant	46.8***	47.7***	45.5***	43.9***
	(3.39)	(3.42)	(3.79)	(3.79)
Grade Fixed Effects	N	N	N	Y
Observations	313	310	310	301
R-squared	0.036	0.045	0.057	0.120

Note: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
This table shows correlates of days attended in the treatment group i.e. lottery-winners who had been offered a Mindspark voucher. Students from poorer backgrounds, and students with higher baseline achievement in Hindi, appear to have greater attendance but the implied magnitudes of these correlations are small. A standard deviation increase in the SES index is associated with a decline in attendance by about 3 days, and a standard deviation increase in Hindi baseline test scores is associated with an additional 5 days of attendance. We find no evidence of differential attendance by gender or by baseline math score.

Table A.4: Quadratic dose-response relationship

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

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

Table A.5: Dose-response of Mindspark attendance

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

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

Table A.6: ITT estimates with inverse probability weighting

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

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

Table A.7: Lee bounds estimates of ITT effects

	(1) Math	(2) Hindi
Lower	0.293	0.162
	(0.084)	(0.092)
Upper	0.434	0.286
	(0.074)	(0.080)
Lower 95% CI	0.153	0.0085
Upper 95% CI	0.557	0.419

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

Table A.8: ITT estimates, by source of test item

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

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

Table A.9: Treatment effect on take-up of other private tutoring

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

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

Appendix B Prior research on hardware and software

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

B.1 Selecting studies

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

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

B.2 Reporting effects

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

In all RCTs, we reported the intent-to-treat (ITT) effect; in all RDDs and IVs, we reported the local average treatment effect (LATE). In all cases, we only reported the magnitude of effect sizes that were statistically significant at the 5% level. These decisions are non-trivial, as the specifications preferred by the authors of some studies (and reported in the abstracts) are only significant at the 10% level or only become significant at the 5% level after the inclusion of multiple covariates. Otherwise, we mentioned that a program had "no effect" on

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

B.3 Categories in each table

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

Table B.1: Impact evaluations of hardware

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

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

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

Table B.2: Impact evaluations of software

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

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

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

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

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

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

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

Appendix C Mindspark software

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

C.1 Computer training

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

C.2 Diagnostic test

After the familiarization routine, students are presented with diagnostic tests in math and Hindi which are used by the Mindspark platform to algorithmically determine their initial achievement level (at which instruction will be targeted). Tests contain four to five questions per grade level in each subject. All students are shown questions from grade 1 up to their grade level. However, if students answer at least 75% of the questions for their corresponding grade level correctly, they can be shown questions up to two grade levels above their own.[35] If they answer 25% or less of the questions for one grade level above their actual grade, the diagnostic test shows no more questions. Initial achievement levels determined by the Mindspark system on the basis of these tests are only used to customize the first set of content that students are provided. Further customization is based on student performance on these content modules and does not depend on their performance on the initial diagnostic test (which is only used for initial calibration of each student's learning level).

C.3 Math and Hindi content

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

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

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

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

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

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

C.4 Personalization

C.4.1 Dynamic adaptation to levels of student achievement

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

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

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

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

C.4.2 Error analysis

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

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

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

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

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

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

C.5 Feedback

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

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

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

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

Figure C.1: Mindspark adaptability in math

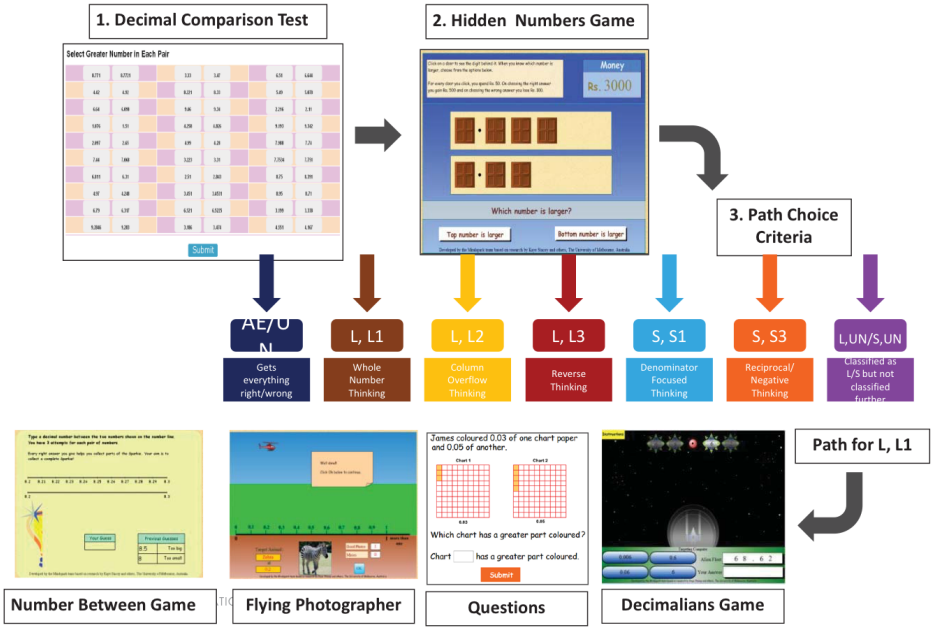
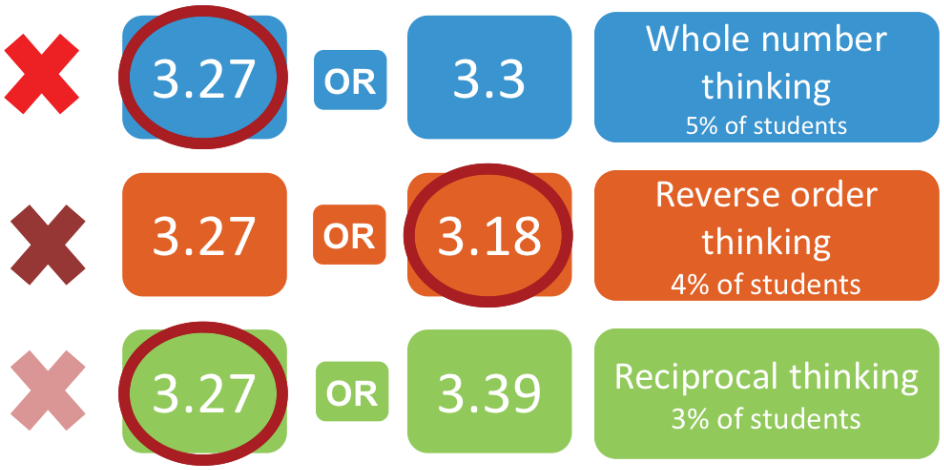


Figure C.2: Student errors in math



Appendix D Test design

D.1 Overview

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

D.2 Objectives of test design

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

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

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

D.3 Test content

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

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

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

D.4 Test booklets

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

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

D.5 Test scoring

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

IRT models specify a relationship between a single underlying latent achievement variable ("ability") and the probability of answering a particular test question ("item") correctly. While standard in the international assessments literature for generating comparative test scores, the use of IRT models is much less prevalent in the economics of education literature

in developing countries (for notable exceptions, see Das and Zajonc 2010, Andrabi et al 2011, Singh 2015). For a detailed introduction to IRT models, please see Van der Linden and Hambleton (1997) and Das and Zajonc (2010).

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

Second, it allows us to robustly link all test scores on a common metric, even with only a partially-overlapping set of test questions, using a set of common items between any two assessments as "anchor" items. This is particularly advantageous when setting tests in samples with possibly large differences in mean achievement (but which have substantial common support in achievement) since it allows for customizing tests to the difficulty level of the particular sample but to still express each individual's test score on a single continuous metric. This is particularly important in our application in enabling us to compute business-as-usual value-added in the control group.[36]

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

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

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

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

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

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

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

D.6 Empirical distribution of test scores

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

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

D.7 Item fit

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

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

Figure D.1: Distribution of raw percentage correct scores

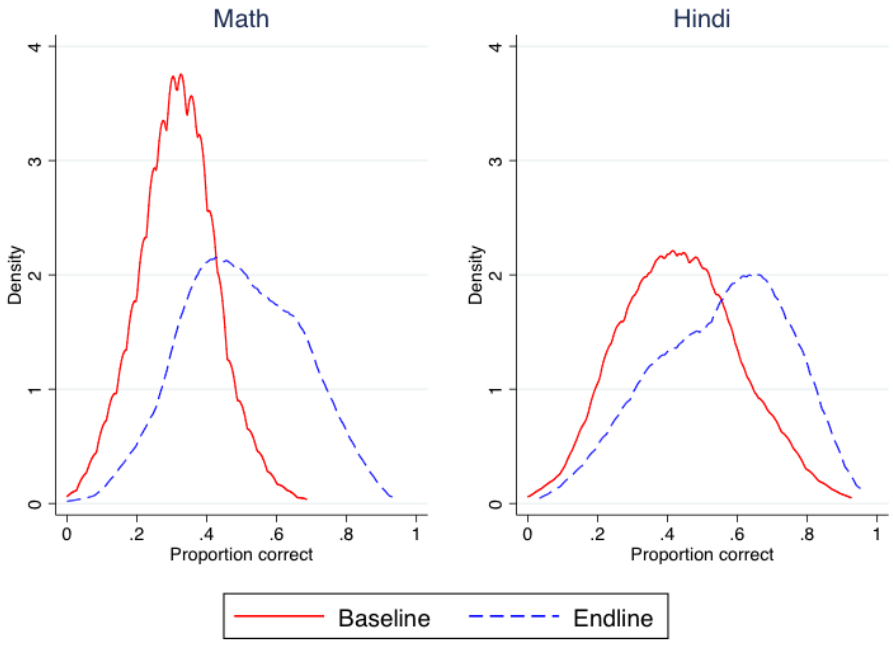


Figure D.2: Distribution of IRT scores, by round and treatment status

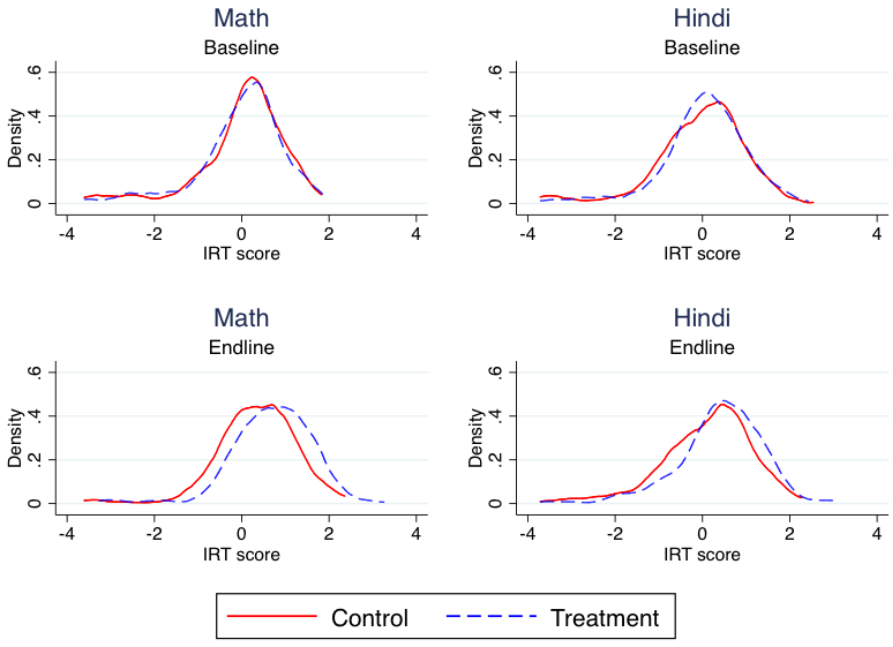
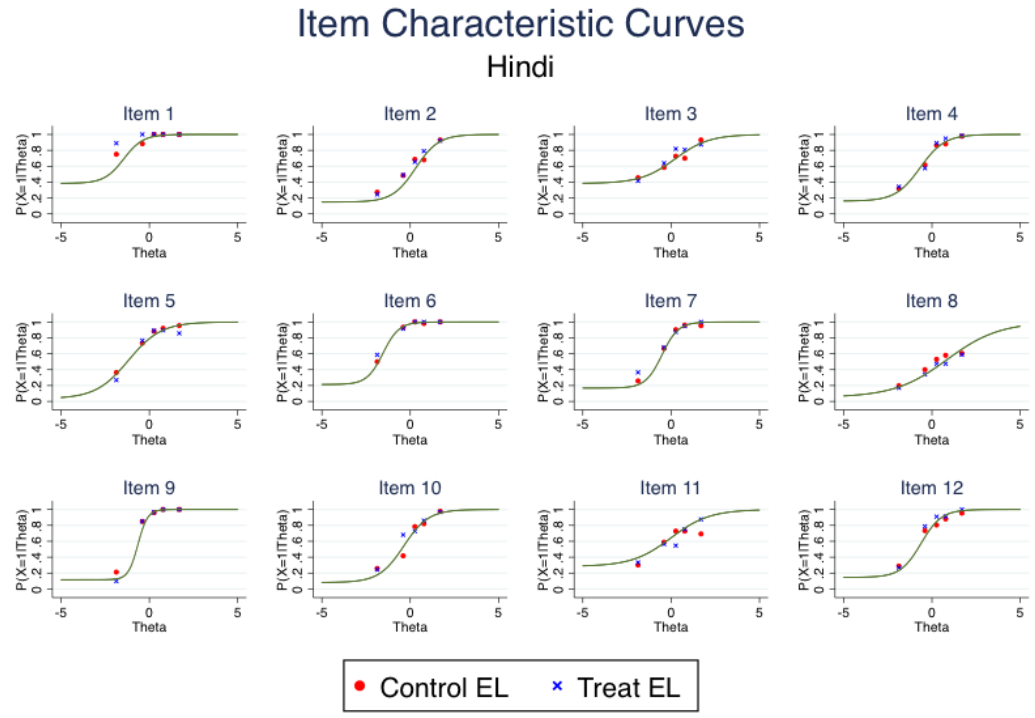
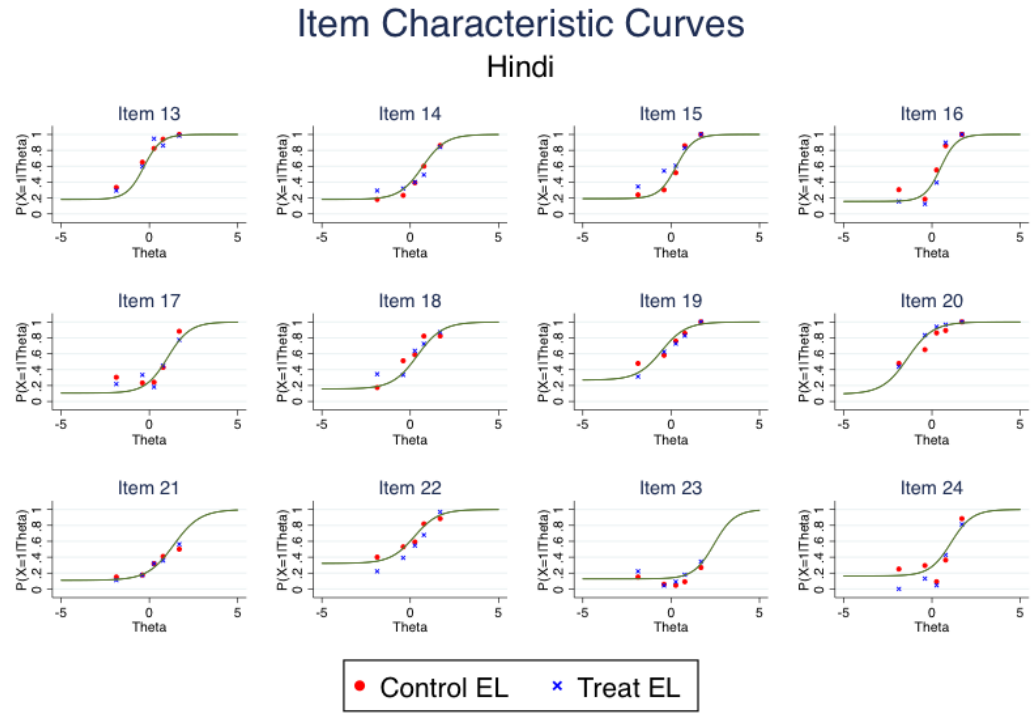


Figure D.3: Item Characteristic Curves: Hindi



Combining all grades



Combining all grades

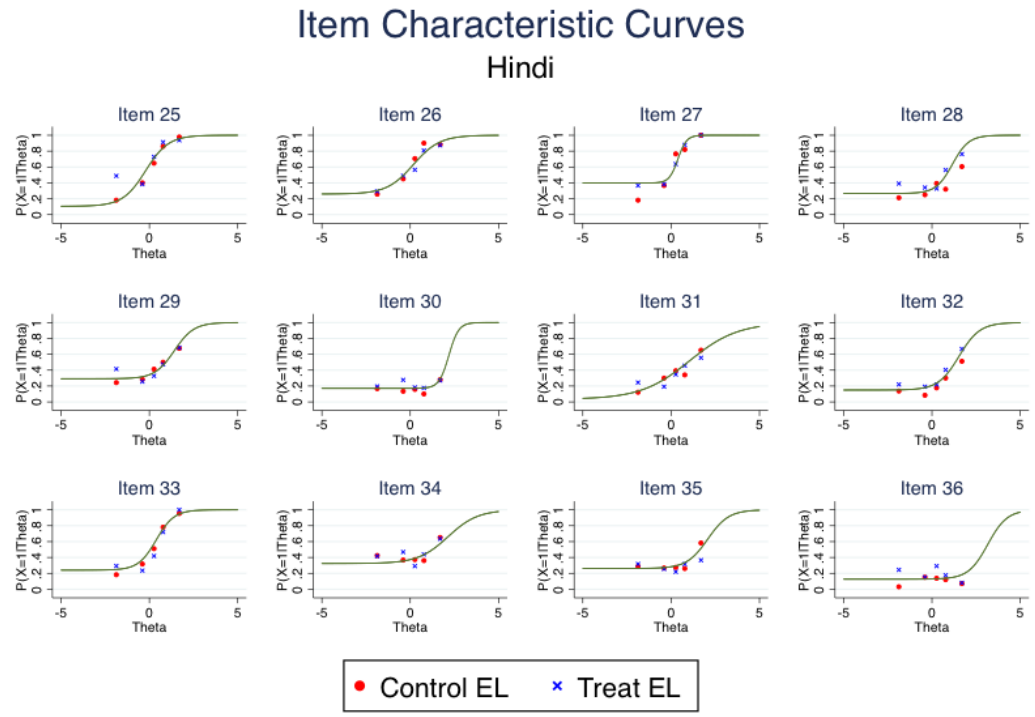
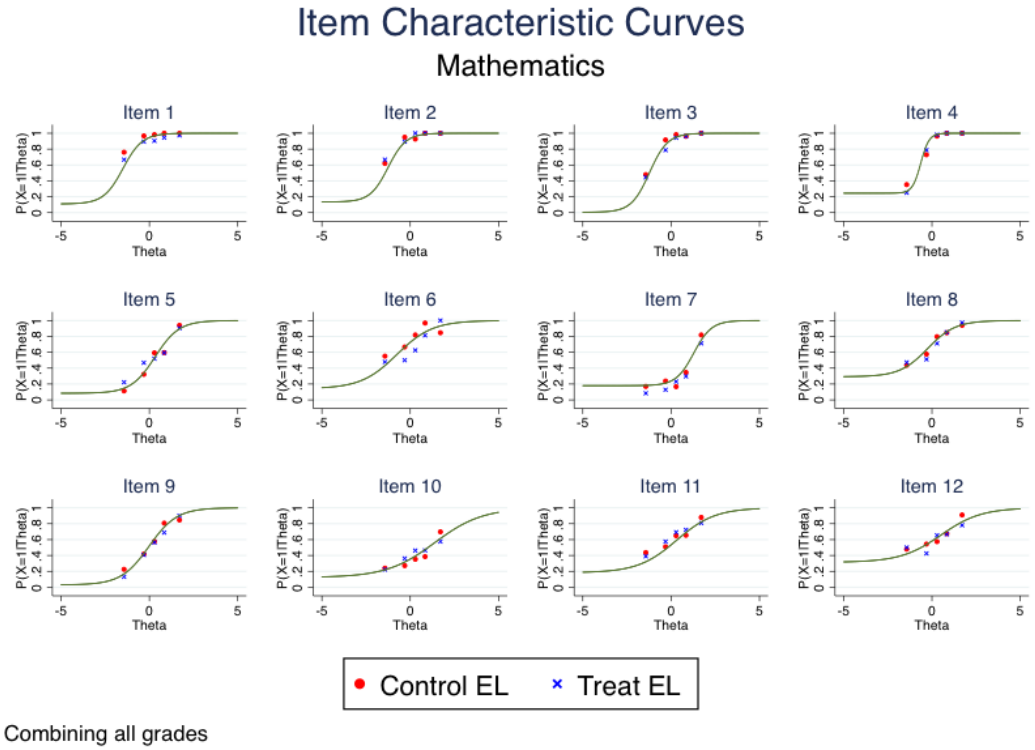
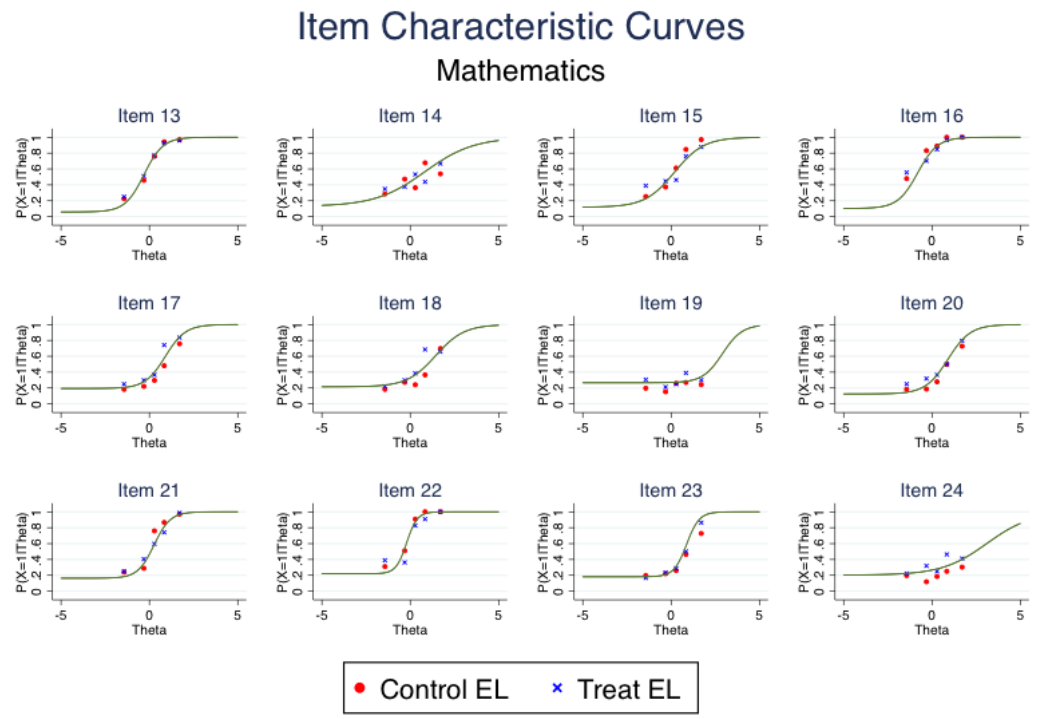
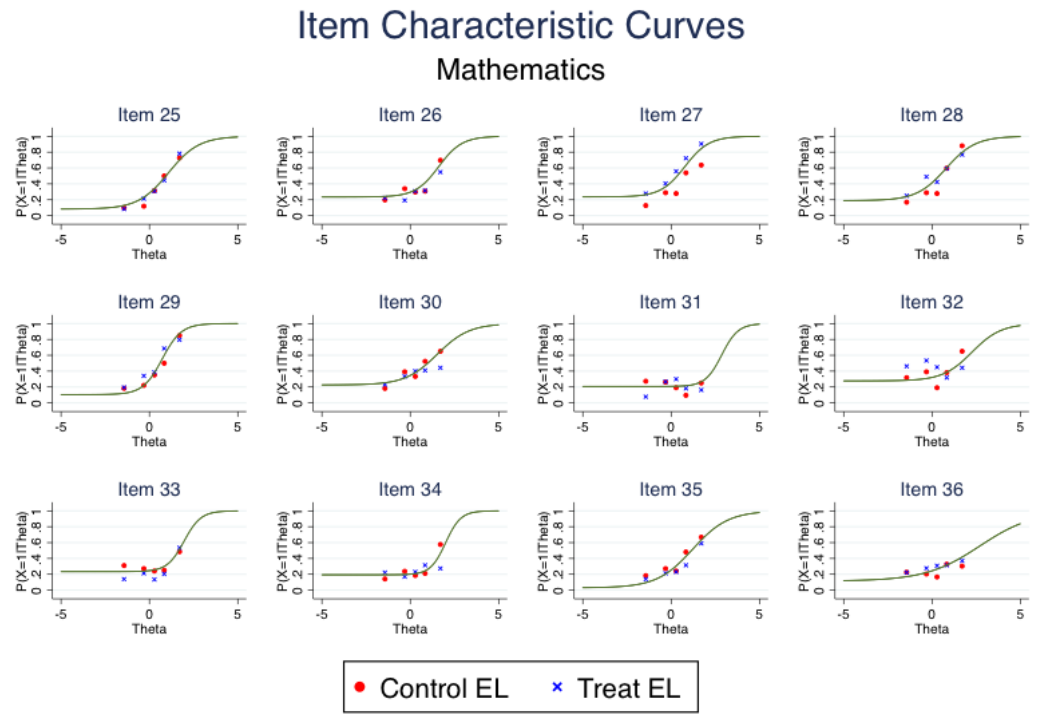


Figure D.4: Item Characteristic Curves: Math





Combining all grades



Combining all grades

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

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

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

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

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

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

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