Which Students Benefit from Personalized Learning? Experimental Evidence from a Math Software in Public Schools in India[*]

Andreas de Barros[†]
Massachusetts Institute of Technology

Alejandro J. Ganimian[‡]
New York University

April 6, 2021

Abstract

This is one of the first studies to evaluate the impact of personalized learning delivered through technology in a developing country. We randomly assigned 1,528 students in grades 6-8 in 15 "model" public schools who were using a computer-adaptive learning software to: (a) a control group, in which they were only able to access the activities for their *enrolled* grade level (i.e., the norm for most software products evaluated in developing countries); or (b) a treatment group, in which they were assigned exercises appropriate for their individual preparation level, across a wide range of grade levels, based on a diagnostic test. After nine months, personalized learning had a null effect on the math achievement of the average student. However, treatment students with low initial performance outperformed their control counterparts by 0.22 standard deviations. Our results suggest that personalized learning is most beneficial for relatively low-performing students, who need help to catch up with their peers.

JEL codes: C93, I21, I22, I25

Keywords: computer-aided learning, India, math instruction, personalized learning

[*] We gratefully acknowledge the funding provided by the Douglas B. Marshall, Jr. Family Foundation for this project. We especially thank Karthik Muralidharan for his involvement as a collaborator in the early stages of this project and for subsequent discussions. We thank Pranav Kothari, Aarthi Muralidharan, Sridhar Rajagopalan, Maulik Shah, Nishchal Shukla, and Gayatri Vaidya for making this study possible. We also thank Anuja Venkatachalam for excellent research assistance and field support. This study was registered with the AEA Trial Registry (RCT ID: AEARCTR-0002459). The usual disclaimers apply. The authors have no conflicting interests to declare.

[†] Postdoctoral Associate, Department of Economics, Massachusetts Institute of Technology. debarros@mit.edu.

[‡] Assistant Professor of Applied Psychology and Economics, New York University Steinhardt School of Culture, Education, and Human Development. alejandro.ganimian@nyu.edu.

There is growing evidence indicating that schoolchildren in many developing countries lag behind their expected grade-level performance, that the gap between expected and actual performance widens during primary school, and that there is wide variation in students' preparation for school within each grade (Andrabi et al. 2007; Das and Zajonc 2010; Duflo et al. 2011; Muralidharan et al. 2019; Pritchett and Beatty 2015). This pattern is expected to be exacerbated by the recent school closures due to the ongoing pandemic (Azevedo et al. 2020; Cummiskey and Stern 2020; Kaffenberger 2020; Kaffenberger and Pritchett 2020).

School systems have sought to address heterogeneity in student preparation for schooling in two main ways: by asking teachers to provide differentiated instruction (also known as "teaching at the right level"; dividing students in groups based on their performance within the classroom and assigning them activities that cater to their level) or computer-adaptive learning (e.g., providing students with access to a software that dynamically adjusts to their level and rate of learning). Differentiated instruction has improved student learning when it is implemented as intended, but teachers have so far been reluctant to integrate this modality into their regular lessons (presumably, because it competes with the pressures they face to complete ambitious curricula and prepare students for high-stakes exams) (see, e.g., Banerjee et al. 2017, 2010, 2007). Computer-adaptive learning (CAL) has also yielded promising results, but the multiple components that CAL software products typically combine make it challenging to understand what, if any, role personalized learning plays in its effectiveness (see, e.g., Banerjee et al. 2007; Muralidharan and Singh 2019; Muralidharan et al. 2019).

This paper presents, to our knowledge, one of the first empirical investigations into the effect of personalized learning, as delivered through a CAL software, on math achievement in a developing country. We randomly assigned 1,528 students in grades 6 to 8 across 15 public schools who had access to a CAL software to learn math to: (a) a control group, in which they were only able to access the activities for their enrolled grade level (i.e., the norm for most software products evaluated in developing countries); or (b) a treatment group, in which they were assigned exercises appropriate for their individual preparation level, based on a diagnostic test. This setup allows to estimate the impact of the personalized-learning feature of the software, which seeks to address heterogeneity in student preparation by allowing students who lag behind curricular expectations to build on their knowledge to reach standards for their enrolled grade. Importantly, we conducted our study in "model" public schools, which serve disadvantaged areas but select students based on an entrance exam (among other differences we discuss below). While similar "exam" schools have received considerable attention in developed countries (e.g., Abdulkadiroğlu et al. 2014; Angrist et al. 2019; Dobbie and Fryer Jr. 2014; Dur et al. 2021; Ellison and Pathak 2021), they have been studied far less in developing countries—except for when they are part of a centralized school-choice system (e.g., Pop-Eleches and Urquiola 2013).

We report three main sets of results. First, as it has been shown in other settings, most students were diagnosed by the software to be several levels behind their enrolled grade (based on a diagnostic test that all students were required to take when they first logged into the software). For example, the average sixth grader was two grade levels behind grade standards in math; for the average eighth-grader, this gap was even larger: two grade levels and a half. Thus, the students in our context, much like those in many low- and middle-income countries, stood to benefit considerably from personalization.

Second, personalized learning had a positive, but statistically insignificant effect on the math achievement of the *average* student in our sample. Relative to the control group, the treatment group performed 0.05 standard deviations (SDs) better in independent assessments of math (not linked to the software), but this difference did not reach conventional levels of statistical significance. Based on our 95% confidence interval, we can rule out negative effects below -0.02 SDs and positive effects above 0.13 SDs.

Third, personalization had a non-trivial positive effect for low-performing students (i.e., those in the bottom quartile of the within-grade baseline math achievement distribution). Initially low-achieving treatment students outperformed their control counterparts by 0.22 SDs. In fact, the intervention had positive and statistically significant effects for this sub-group in two of the three content domains and cognitive domains assessed at endline, suggesting these gains were broad based. We did not, however, find evidence of heterogeneous effects along other dimensions, such as students' sex and grade.

Our study makes three main contributions to existing research. First, it is one of the first to examine the effect of a *specific feature* of a highly effective CAL software (in a similar fashion to de Barros et al. (2020)). This is a crucial contribution because prior studies have mostly evaluated "bundled" interventions that combine multiple features, which makes it challenging for researchers and policymakers to understand what makes an effective education-technology intervention (for a discussion of the challenges involved in drawing conclusions from bundled interventions in education, see Muralidharan (2017)). It is also important because the software used in our study (called Mindspark and developed by a local leading assessment firm, Educational Initiatives) has proven to be one of the most effective ones to have been evaluated in a developing country (see Muralidharan and Singh 2019; Muralidharan et al. 2019).

Second, our study is one of the first to illustrate the benefits of personalized learning when delivered through technology (instead of through teacher-led differentiated instruction). This is particularly important because, while differentiated instruction has proven effective when delivered during summer camps (when school is not in session), before or after school, and during the school day with the assistance of additional instructors, teachers remain reluctant to implement it during the school day when they are on their own (see Banerjee et al. 2011).

Therefore, our results indicate that when schools have the infrastructure and requisite teaching staff to adopt a CAL software, personalization via technology can be an effective policy alternative to improve the learning outcomes of the lowest-performing students.

Third, and more broadly, our study contributes to growing evidence that the potential of education-technology interventions to complement traditional (i.e., "chalk-and-talk") instruction depends largely on the baseline quality of instruction. Nearly a decade ago, Linden (2008) found that the same CAL software that had previously had positive effects in schools serving disadvantaged children in Gujarat, India had negative impacts in a well-functioning network of schools run by a non-profit in the same state. Ever since, multiple studies have sought to distinguish between the effect of the content of CAL software and the additional inputs with which it is usually delivered (see, e.g., Bettinger et al. 2020; Ferman et al. 2019; Ma et al. 2020; Mo et al. 2020), each with its own set of challenges (discussed in Ganimian et al. 2020). Our results offer further evidence that the margin for impact of technology-enabled personalization may be limited for the average student attending relatively well-functioning public schools.

The rest of the paper is structured as follows. Section 1 context, study design, and intervention. Section 2 describes the data. Section 3 discusses the empirical strategy. Section 4 reports the results. Section 5 discusses implications for research and policy.

1 Experiment

1.1 Context

Schooling in India is compulsory and free from ages 6 to 14 (Ministry of Law and Justice 2009). Primary education runs from grades 1 to 5 and upper primary runs from grades 6 to 8; grades 1 to 8 are collectively referred to as "elementary education". The Indian school system included 840,241 primary schools, 287,265 upper-primary schools, and 48,543 primary schools with secondary grades in the 2016-2017 school year (NIEPA 2018). That same year, government (i.e., public) schools served nearly a third of students in elementary grades (111,310,953 students, or 59% of total enrollment).

We conducted this study in partnership with Educational Initiatives (EI), a leading assessment firm in the country that developed the CAL software that we used to randomly assign students to personalized learning (described in greater detail in the Intervention sub-section). We established this partnership as a multi-year project to leverage both the vast item bank of the CAL software in math and other subjects and its high degree of penetration across the country to use randomized experiments to answer questions of import to educators. The partnership, dubbed the Learning Lab, was led by Karthik Muralidharan at the University of California, San Diego and Sridhar Rajagopalan at EI and funded by the Douglas B. Marshall, Jr. Family Foundation. We were co-principal investigators on this project.

We conducted this study in the state of Rajasthan, which is an ideal setting to understand the effect of interventions that could be scaled to the rest of India. First, it represents a sizeable share of the country's land and population: it is the largest state in terms of area and the seventh-largest in terms of population (MHA 2012). Second, it is a mostly rural state, much like the rest of the country: three-fourths of its inhabitants live in rural areas. Third, its education level is very similar to that of the rest of rural India: in 2018, only 25% of sixth-graders could solve a subtraction of a two-digit number by another two-digit number and just 29% could perform a division of a three-digit number by a two-digit number, compared to 24% and 35% of sixth-graders in all of rural India (ASER 2019).

Specifically, we conducted our study in "model" public schools. These schools were created in 2009 by the Ministry of Human Resource Development at the central government to promote education in rural areas. They differ from regular public schools in five main ways: they focus on "educational backwards" areas of the state (as opposed to the entire state),[1] they only cover grades 6 to 12 (instead of grades 1 to 12), English is their medium of instruction (rather than the local language), they follow a curriculum prescribed by the national government (instead of the one set by the state government), and they require that students reach a minimum score on an entrance exam to gain admission (Kumar 2020).[2] In 2017, the year of our study, there were 134 model schools across Rajasthan; in fact, 303 (64%) of all "blocks" (i.e., district sub-divisions) in the state had at least one model school. These schools were particularly well suited for our study because they are required to meet requirements for infrastructure (e.g., running electricity, Internet connectivity, and computer laboratories) and teaching (e.g., full-time computer-science teachers and regular weekly slots for computer-science lessons) that make it easier to deploy an educational software.

1.2 Sample

The sample for the study included 1,528 students from grades 6 to 8 across 15 public model schools across seven districts in Rajasthan: Alwar, Bhilwara, Bundi, Dungarpur, Jodhpur,

[1] Educational backwards areas are those in which the rural female literacy rate is below the national average and the gender gap in literacy is above the national average.

[2] Model public schools follow the curriculum of the Central Board of Secondary Education (CBSE), set by the National Council of Educational Research and Training (NCERT), whereas regular public schools follow the curriculum of the Rajasthan State Board of Secondary Education (SSE), set by the state government. In practice, however, the SSE draws heavily on the CBSE. There are three main differences between these curricula: English is a primary language in CBSE, but a secondary language in SSE; CBSE is more widely recognized across India; and all national examinations for entry into top colleges in India are based on the CBSE.

Rajsamand, and Udaipur (see Figure A.1 in Appendix A). We drew a convenience sample of schools based on three criteria: (a) they had fully constructed buildings; (b) they had space for a computer lab; and (c) they had running electricity.[3] All 15 schools agreed to participate. We sought informed consent from principals and teachers at those schools.

Attrition from the study was non-trivial: 1,078 (or 71%) of the 1,528 students who participated in the baseline assessments also took the endline assessments. Yet, we find no evidence of differential attrition by experimental group: 28% of control students and 31% of treatment students who were present at baseline missed the endline, and the difference is not statistically significant. To verify that our general pattern of results are not affected by attrition, we include both inverse-probability weighted (IPW) estimates and Lee (2009) bounds.

1.3 Randomization

We randomly assigned the 1,528 students in our sample to: (a) a control group, in which students were only able to access the activities in a computer-assisted learning (CAL) software for their *enrolled* grade level (762 students); or (b) a treatment group, in which students were assigned exercises appropriate for their individual preparation level, across a wide range of grade levels, based on a diagnostic test (766 students). We describe the differences in the experience of each group in the Intervention section below.[4]

We stratified the randomization by grade-by-section combination to maximize statistical power. One potential drawback of this approach is that it allows for spillovers across students in the same classroom (e.g., if students T and C are in the same classroom, and T is assigned to the treatment group and C to the control group, if T learns more due to the intervention, he/she may teach some of what he/she learned to C, causing him/her to learn more in spite of not having received the intervention). Yet, given that the benefit from the contrast between groups stems from personalization, we think this is unlikely to be a major concern. Additionally, principals, teachers, and students were "blind" to the experimental condition to which students were assigned, which further reduced the likelihood of spillover effects.

Control and treatment students were comparable on their baseline achievement and sex, regardless of whether we consider all students present at baseline or only those who also took the endline assessment (i.e., non-attritors, see Table 1). In fact, not just the means, but the distribution of baseline achievement was quite similar across experimental groups (see Figure A.2).

[3] Model schools are required by the state government to meet all three of these requirements (see previous sub-section), but in practice, this is not always the case.

[4] We initially planned to introduce two more treatment groups. Yet, due to technical difficulties, we abandoned these two interventions after three months and excluded students in those experimental groups from this study.

1.4 Intervention

We provided all students in our study with a CAL software called "Mindspark", which focused on math instruction.[5] The software was developed by Educational Initiatives (EI), a leading assessment firm in India, over a 10-year period. It has been used by over 500,000 students, it has a database of over 45,000 questions, and it administers over 2 million questions across its users every day. It can be delivered during the school day, before or after school at stand-alone centers, and through a self-guided online platform. The after-school version was recently evaluated through a randomized experiment and found to vastly improve the math and reading achievement of primary and middle-school students in Delhi (Muralidharan et al. 2019). The in-school version, which is the one that we use in the present study, is currently being evaluated in Rajasthan. Its impacts are smaller than those of the after-school version, but they are commensurate with the lower dosage that students receive in this model, which are also achieved at lower costs (Muralidharan and Singh 2019).

In the present study, we are not interested in evaluating the impact of the software; instead, we use it to estimate the effect of its personalization feature on students' math achievement. The software works as follows. When students first log in, they are asked to take a brief diagnostic test, which identifies what they know and are able to do, and the areas in which they can improve. This test also determines the grade level at which the student can answer most questions, which may or may not be his/her enrolled grade (e.g., a student may be *enrolled* in grade 6, but *perform* at a grade-4 level). Then, the software presents the student with a number of exercises on topics appropriate for their preparation level, based on the diagnostic test. The difficulty and topic covered by subsequent exercises dynamically adjusts to the students' progress (e.g., a student who answers most exercises correctly may be presented with more difficult exercises, whereas a student who answers exercises incorrectly may be presented with easier exercises, and he/she may even be redirected to remedial exercises). In this study, we temporarily restricted the exercises that the control students could access to those associated with their enrolled grade level. Students interacted with the software in their computer labs, with the assistance of a "lab in-charge", who opened and maintained the computer labs (i.e., not their math teacher). This setup allows us to estimate the effect of students being able to access exercises more closely aligned with their preparation.

Importantly, the version of the CAL software that the control students were offered resembles most educational software products that have been evaluated in developing countries, which are used to allow students to practice what they learn at school on a given week and thus focuses on the content prescribed for their enrolled grade level (see, e.g., Carrillo et al. 2011; Lai et al. 2015, 2013, 2012; Mo et al. 2020, 2014, 2015). The control version of the software does

[5] The software can also provide language instruction, but this function was deactivated in our study.

allow for some degree of personalization within grade-appropriate materials, but students are not presented with material for lower grades regardless of how far behind they lag in their performance.[6] This design feature of our study allows us to shed light on the potential contribution of the personalization feature to computer-aided learning.

A few of the educational software products that have been evaluated in developing countries include some degree of personalization (e.g., Banerjee et al. 2007; Linden 2008). Yet, the vast item bank and learning pathways of the CAL software that we use in this study provide a greater degree of personalization of both the content and difficulty of the material.

2 Data

We collected two main types of data: (a) students' achievement, before and after the intervention, to check for baseline equivalence and estimate impact; and (b) students' usage of the CAL software and interaction with the intervention, to verify implementation fidelity. We complemented these data with administrative information on students' grade and sex (we did not, however, conduct a student survey).

2.1 Student achievement

We administered student assessments of math at baseline (before the intervention) and endline (37 weeks after the start of the intervention).[7] These assessments evaluated what students ought to know and be able to do according to international standards, including three content domains (numbers, geometric shapes and measurement, and data visualization) and three cognitive domains (knowing, applying, and reasoning). The distribution of items across content and cognitive domains was based on the assessment framework of the 2019 Trends in International Math and Science Study (TIMSS) for grade 4 (Mullis and Martin 2017).

Each test had 35 multiple-choice items. We drew on items from international assessments (e.g., TIMSS, PISA, Young Lives), domestic assessments (e.g., Quality Education Study, Student Learning Survey), and previous impact evaluations in India (e.g., the Andhra Pradesh Randomized Studies in Education or APRESt). We included items from a wide range of

[6] For example, if a control group student gets introduced to fractions and struggles, she may be asked to slow down and review additional materials, at grade level. However, he/she would not be exposed to remedial materials from lower grades, such as learning units that focus on basic number sense (a potential prerequisite to learn fractions).

[7] Different schools conducted the baseline assessment on slightly different dates, from August 28 to September 2, 2017. They also conducted the endline assessment on different dates, from April 8 to 25, 2018. However, the software was activated on August 6, 2017 for 97% of study participants, so all our analyses focus on the period between that date and April 20, 2018 (which amounts to 37 weeks).

difficulty to reduce the possibility of students not answering any questions correctly and students answering all questions correctly. We designed a single assessment for all grades in the study (i.e., grades 6 to 8) in each round (i.e., baseline and endline), but we created three versions of the assessment at baseline and four versions at endline to prevent students from cheating.[8]

We used a non-equivalent anchor test (NEAT) design to link results across administrations (for a discussion of this design, see Kolen and Brennan 2004). We included an "anchor test" with overlapping items across rounds of data collection and we scaled the results for both rounds concurrently using a two-parameter logistic Item Response Theory (IRT) model.

Importantly, the baseline assessments were administered roughly two weeks *after* the software was activated in study schools, so in theory, students' baseline scores could reflect what students learned by using Mindspark during those two initial weeks. In practice, however, the average student was exposed to the software for only 21 minutes during this period, so we think it is unlikely that it produced any meaningful changes in student achievement in math. (We discuss students' exposure to the program during the study in greater detail in the Results section). Further, as we show below, our impact estimates remain virtually unchanged when we do not account for baseline performance, suggesting that this is unlikely to be a major concern.

2.2 Students' interaction with the software

We also obtained data on students' interaction with the CAL software. These include: (a) students' initial preparation (from the diagnostic test, described in the Intervention sub-section); (b) the time that students spent interacting with the software during each session (from a unique identifier that students use when they log into the CAL platform); (c) the difficulty level of the exercises to which they were presented (benchmarked against expected performance in each grade by the software developers); (d) the time it took each student to attempt each exercise; and (e) whether he/she answered each exercise correctly.

3 Empirical strategy

We estimate the effect of the offer of personalization (i.e., the intent-to-treat or ITT effect) by fitting the following model:

$$Y_{igs}^{t} = \alpha_{r(gs)} + \beta T_{igs} + \gamma Y_{igs}^{t-1} + \epsilon_{igs}^{t} \tag{1}$$

[8] The tests can be accessed at https://bit.ly/3knzgaj (baseline) and https://bit.ly/2E8U0mF (endline).

where Y^t_{igs} is the math achievement of student i in grade-by-section g and school s at time t (endline), $r(gs)$ is the randomization stratum of grade-by-section g and school s and $\alpha_{r(gs)}$ are stratum fixed effects, T_{igs} is an indicator variable for random assignment to treatment, and Y^{t-1}_{igs} is the math achievement of the student at $t-1$ (baseline). The parameter of interest is β, which captures the causal effect of the intervention. We fit variations of this model that interact the treatment dummy with students' grade, sex, and baseline achievement (continuous or by within-grade quartile) to understand whether the intervention is more helpful for some sub-groups of students. We also interact the treatment dummy with the three student characteristics that we observe at baseline (i.e., sex, grade, and initial performance) to test for heterogeneous effects. Finally, as mentioned in the previous section, we also use IPW estimates and Lee (2009) bounds to show that attrition does not alter our general pattern of results.

4 Results

4.1 Implementation fidelity

The intervention was implemented largely as intended. First, virtually all students across both experimental groups (1,069 out of 1,078 students or 99.2%) logged in at least once to the CAL platform during the evaluation. The total time that students spent interacting with the software, however, was relatively low: the typical student (i.e., in the 50th percentile of the usage distribution) interacted with the CAL software for 329 minutes during the nine months of the intervention (Figure 1). This level of exposure is considerably lower than that of the out-of-school version of the program evaluated in Delhi (Muralidharan et al. 2019),[9] but it reflects the constraints that schools face to integrate this software into their regular instruction (e.g., availability of classrooms and equipped computers, coordination between teachers' timetables, time lost by taking students to the computer lab) (see, e.g., Ferman et al. 2019; Rodriguez-Segura 2020).

Exposure to the software varied widely across students: the least frequent users (i.e., those in the 25th percentile of the usage distribution) interacted with the software for less than 250 minutes during the study, whereas the most frequent users (i.e., those in the 75th percentile of the distribution) had twice as much exposure, totaling nearly 500 minutes in the same period. Exposure also varied over time: in some weeks, no student had any interaction with the software, whereas in others usage was up to 30 minutes. This variation suggests that our results should be interpreted as lower bound estimates of the effects of personalization on math

[9] If the median student in our study had interacted with the CAL software with the same intensity as his/her counterparts in the evaluation of the out-of-school version of the program in Delhi, he/she would have spent 1,260 minutes using the software throughout the intervention period.

achievement, which could be improved upon if schools were able to increase and sustain use of the software.

For students who were exposed to the software, the diagnostic assessment confirmed that they had a clear need for personalized learning. We observed two patterns documented in previous studies. First, the average student lagged far behind curricular expectations for his/her grade: for example, the average student enrolled in grade 6 actually performed at a grade 4 level in math (Figure 2). Second, there was wide variability in student achievement within each grade: for example, while some grade 6 students performed at a grade 2 level, others performed at a grade 7 level (Figure 2). No teacher, no matter how effective, could possibly provide personalized instruction to students at such disparate levels of preparation, so the CAL software was, in theory, well positioned to complement teacher-led instruction.

The randomization of the personalization feature worked exactly as expected. Control students were presented with exercises that corresponded to their *enrolled* grade (e.g., grade 6 students only saw grade 6 exercises), whereas treatment students were offered exercises that corresponded to their *diagnosed* grade (e.g., grade 6 students only saw grade 6 exercises), whereas treatment students were offered exercises that corresponded to their diagnosed grade (e.g., grade 6 students diagnosed to be at a grade 3 level saw grade 3 exercises, see Figure 3). Also, while control students continued to be presented with exercises matching their enrolled grade level, their treatment counterparts saw increasingly more difficult exercises during the study period (e.g., grade 7 students started attempting exercises at a grade 4 level and, by the end of the experiment, they were completing exercises between grades 5 and 6, see Figure 4). Similarly, while control students attempted exercises matched to their enrolled grade level regardless of their initial diagnostic, their treatment peers started at their diagnosed level and "graduated" to higher levels (e.g., students diagnosed to be at grade 7 started attempting exercises at a grade 5 level; by the end of the experiment, they were completing exercises at a grade 7 level; see Figure 5). In other words, as stated above, the software not only matched students' initial *level* of preparation, but also their *rate of progress* (i.e., increasing difficulty more rapidly for students who answered more questions correctly).

The exercises attempted by study participants during the evaluation focused mostly on numbers (95% of the total), and much less on geometry (4.5%) or data (0.5%, Table A.1).[10] Specifically, the most featured topics were: whole-number concepts (about 28% of the total), whole-number operations (19%), real numbers (15%), integers (9%), number theory (8%) and basic algebra (7%, see Table A.1). As we argue below, this distribution of exercises is helpful to understand why students' achievement improved in some topics and not others.

[10] The mapping of exercises to topics was conducted by Educational Initiatives, the developer of the CAL software, prior to the start of the study. The grouping of topics into content domains was conducted by us at the analysis stage.

4.2 Average effects on math achievement

The offer of the intervention had a null effect (of about 0.05 SDs) on the math achievement of the average student, regardless of whether we account for students' performance at baseline on the assessments we developed and administered or on the software's own diagnostic test (Table 2). In fact, based on the 95% confidence interval, we could rule out effects below -0.02 SDs and above 0.13 SDs. When we estimated effects separately by content and cognitive domain, we observed effects for data-related items and items in which students were asked to apply their knowledge (of about 0.02 pp. in both cases; Table 3, panel A). Yet, both of these effects are small (below 2.3 pp. in both cases) and neither effect is statistically significant once we account for multiple hypothesis testing with a family-wise error rate p-value adjustment (following List et al. 2019). Lastly, while the effect of personalization varied across schools, the differences across schools were not statistically significant in any case (Figure A.3. Together, these results suggest that the average student benefited little from personalization.[11]

We found no evidence that the average effects of the intervention were affected by student attrition from baseline to endline. In Table 2, column 2 shows that our estimate of the average intent-to-treat (ITT) effect remained virtually unchanged if we weighted results by the inverse probability of each student of participating in the endline. Further, when we estimated Lee (2009) bounds, the lower and upper bounds of the treatment effects were both positive, but we could not reject the null hypothesis that the lower bound was equal to zero (see Table A.2).

4.3 Heterogeneous effects on math achievement

The null average effects, however, masked important heterogeneous impacts. We investigated whether the effect of personalization differed across three pre-specified student characteristics recorded in our data: sex, enrolled grade, and initial achievement. Notably, we found that the intervention had a medium-to-large positive effect of 0.22 SDs for students with initially low achievement in math (i.e., those in the bottom quartile of the within-grade baseline math achievement distribution). We first show this graphically, by plotting the treatment effects by students' baseline quartile (Figure 6) and we then demonstrate this analytically in two ways: by accounting for students' baseline performance and interacting it with the treatment indicator, and by interacting the treatment indicator with indicators for each student's within-grade quartile of baseline achievement (Table 4, columns 1 and 2). These findings maintain their statistical significance at the 10% level, after accounting for multiple hypothesis testing. We also observe this pattern when we plot effects by students' baseline performance (Figure A.4).

[11] We do find a statistically significant effect of personalization on items that were first introduced in the endline (Table A.3, panel A). However, given all other results, we believe that this is likely to have occurred by chance.

To investigate whether those students who were diagnosed to be lagging behind improved more, we examined heterogeneous effects by students' performance on the diagnostic assessment administered by the software upon students' first login (see Intervention sub-section). We interacted the treatment with students' within-grade percentile on the diagnostic test (Table 4, columns 3), and with the difference between each student's enrolled and diagnosed grade level (Table 4, column 4). Both specifications indicated that students with lower math achievement at baseline saw larger treatment effects than their peers (as suggested by the coefficients on the treatment indicator and the respective interaction terms). However, neither the treatment effects on the lowest-performing students nor the interactions remained statistically significant at the 10% level after accounting for multiple hypothesis testing.

The improvements made by low-performing students were concentrated in one content domain (numbers) and one cognitive domain (applying knowledge; see Table 3, panel B). This pattern is not surprising, given that (as we stated in the sub-section on implementation fidelity), the vast majority of the exercises attempted by study participants focused on numbers (see Table A.1). Once we account for multiple-hypothesis testing, however, only the impact on applying knowledge retains statistical significance.

Importantly, the effects on low-performing students were not merely a result of "teaching to the test." These students improved their performance both on items that were administered in baseline and endline by 2.8 pp. (which we call "repeated items") and on items that were first introduced in the endline by 6.8 pp. (which we call "non-repeated items"; see Table A.3).

We did not find any evidence of heterogeneous effects by students' sex: female students performed slightly below male students (by 0.03 SDs), but the difference was not statistically significant, nor was the interaction between the treatment and female indicator (Table A.4). We did not find any evidence of heterogeneity in treatment effects by students' enrolled grade.

5 Conclusion

This paper presents one of the first studies that isolates the effect of technology-enabled personalized learning in a developing-country setting. After about nine months, we found that students who could access exercises that were below or above their enrolled grade level performed, on average, no differently from those who were only allowed to access exercises at their enrolled grade level. However, low-performing students (i.e., those who performed in the lowest quartile of the baseline math achievement distribution) in the treatment group outperformed their control counterparts by 0.22 SDs. Reassuringly, these gains were concentrated in the topics and skills that were featured more frequently in the software. Yet, they did not reflect "teaching to the test," given that they affected items administered at baseline and endline as well as new items. This pattern of results suggests that personalized learning matters most to low-performing students, who arguably get very little from exercises that focused on material for their enrolled grade, given that they perform several grade levels behind curricular expectations for their grade.

Our study makes several important contributions. First, it adds to our ongoing understanding of why computer-adaptive learning (CAL) software products may be among the most effective education-technology interventions evaluated in developing countries to date (see Ganimian et al. 2020). Specifically, our study suggests that the personalization feature in the Mindspark software, found effective in both its after-school (Muralidharan et al. 2019) and in-school formats (Muralidharan and Singh 2019), may play an important part in improving students' learning outcomes. This finding is intuitive, and arguably even obvious after the fact, but to our knowledge, there are no experimental studies in developing countries that are designed to isolate the effects of personalization from other features of CAL software products. The vast majority of impact evaluations of education technology interventions in developing countries focus on estimating the effect of multifaceted software products, which include personalization as well as many other features (for reviews, see Bulman and Fairlie 2016; Escueta et al. 2020; Tauson and Stannard 2018). The present study is a much-needed step in identifying the *features* that make some products effective—especially, for low performers.

Second, our study also contributes to the growing evidence of personalization broadly, even when it is not delivered through technology. Over the past two decades, a number of impact evaluations have found that, when teachers are able to teach to a more homogenous student group—e.g., through ability tracking (Duflo et al. 2011) or differentiated instruction within the classroom (Banerjee et al. 2010, 2007)—they improve the achievement of their students by a greater margin than when they teach all of their students at once. There is a growing consensus in development policy circles that this might be because, as we find in our study, many students in developing countries perform well below curricular expectations for their grade, and thus stand to benefit from reviewing below-grade-level skills. Our study not only provides evidentiary support for that hypothesis, but also sheds light on the comparative advantage of technology to provide such personalization.

Third, our study highlights the importance of the counterfactual (i.e., regular-instruction) conditions in evaluations of technology-enabled interventions. Specifically, it suggests that such interventions have a relatively narrow margin to impact the average student in settings where the baseline quality of teacher-led instruction is better than in the typical public school, as it may be the case with the model public schools in our study. Yet, equally importantly, these interventions may still improve the performance of low achievers, who may not have reached a performance level that allows them to reap the benefits of better-than-average instruction.

Finally, our study demonstrates how to leverage the increasing prevalence of educational software products to run rapid-cycle randomized evaluations that shed light into the merits of intuitively appealing yet largely untested educational strategies. Perhaps more importantly, it does so in a way that allows researchers to closely monitor students' interaction with the intervention being tested (in this case, personalized instruction) and to estimate its effect not just for the average student, but also for relevant sub-groups. We see this as a crucial contribution to research on education technology, given that many interventions that have been evaluated in this space have yielded disappointing results and would benefit from feedback on their effectiveness (Ganimian et al. 2020).

References

Abdulkadiroğlu, A., J. Angrist, and P. Pathak (2014). The Elite Illusion: Achievement Effects at Boston and New York Exam Schools. *Econometrica 82*(1), 137–196. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA10266.

Andrabi, T., A. I. Khwaja, T. Vishwanath, and T. Zajonc (2007). Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate. *Unpublished manuscript*. Washington, DC: The World Bank.

Angrist, J. D., P. A. Pathak, and R. A. Zárate (2019, August). Choice and Consequence: Assessing Mismatch at Chicago Exam Schools. Technical Report w26137, National Bureau of Economic Research.

ASER (2019). Annual Status of Education Report 2018 (Rural). Provisional Report, Pratham, New Delhi.

Azevedo, J. P., A. Hasan, D. Goldemberg, S. A. Iqbal, and K. Geven (2020, June). Simulating the Potential Impacts of COVID-19 School Closures on Schooling and Learning Outcomes: A Set of Global Estimates. Working Paper 9284, The World Bank, Washington, D.C.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017, November). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010, February). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy 2*(1), 1–30.

Banerjee, A. V., R. Banerji, E. Duflo, and M. Walton (2011). Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India. *Unpublished manuscript*. New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL).

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

Bettinger, E., R. Fairlie, A. Kapuza, E. Kardanova, P. Loyalka, and A. Zakharov (2020, April). Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function. Technical Report w26967, National Bureau of Economic Research, Cambridge, MA.

Bulman, G. and R. W. Fairlie (2016). Technology and education: Computers, software, and the internet. Technical report, National Bureau of Economic Research.

Carrillo, P. E., M. Onofa, and J. Ponce (2011). Information technology and student achievement: Evidence from a randomized experiment in ecuador.

Cummiskey, C. and J. Stern (2020). Calculating the Educational Impact of COVID-19 (Part II): Using Data from Successive Grades to Estimate Learning Loss.

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

de Barros, A., A. J. Ganimian, and A. Venkatachalam (2020, August). How Much Do Students Benefit from Practice Exercises? Experimental Evidence from India.

Dobbie, W. and R. G. Fryer Jr. (2014, July). The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools. *American Economic Journal: Applied Economics 6*(3), 58–75.

Duflo, E., P. Dupas, and M. Kremer (2011, August). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review 101*(5), 1739–1774.

Dur, U., P. Pathak, and T. Sönmez (2021). Explicit vs. Statistical Preferential Treatment in Affirmative Action: Theory and Evidence from Chicago's Exam Schools. *Journal of Economic Theory Forthcoming*.

Ellison, G. and P. A. Pathak (2021, March). The Efficiency of Race-Neutral Alternatives to Race-Based Affirmative Action: Evidence from Chicago's Exam Schools. *American Economic Review 111*(3), 943–975.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020, December). Upgrading Education with Technology: Insights from Experimental Research. *Journal of Economic Literature 58*(4), 897–996.

Ferman, B., L. Finamor, and L. Lima (2019, June). Are Public Schools Ready to Integrate Math Classes with Khan Academy? Working Paper 94736, University of Munich, Munich.

Ganimian, A. J., F. M. Hess, and E. Vegas (2020). Realizing the promise: How can education technology improve learning for all? Technical report, Brookings Institution, Washington, D.C.

Kaffenberger, M. (2020, June). Modeling the Long-Run Learning Impact of the COVID-19 Learning Shock: Actions to (More Than) Mitigate Loss. Working Paper, Research on Improving Systems of Education (RISE), Oxford, UK.

Kaffenberger, M. and L. Pritchett (2020, May). Failing to Plan? Estimating the Impact of Achieving Schooling Goals on Cohort Learning. Working Paper 20/038, Research on Improving Systems of Education (RISE), Oxford, UK.

Kolen, M. J. and R. L. Brennan (2004). *Test Equating, Scaling, and Linking* (3rd ed.). New York, NY: Springer.

Kumar, N. (2020, July). Public School Quality and Student Outcomes: Evidence from Model Schools in India. Publication Title: 2019 Papers.

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

Lai, F., L. Zhang, X. Hu, Q. Qu, Y. Shi, Y. Qiao, M. Boswell, and S. Rozelle (2013). Computer assisted learning as extracurricular tutor? evidence from a randomised experiment in rural boarding schools in shaanxi. *Journal of Development Effectiveness 5*(2), 208–231.

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

Lee, D. S. (2009, July). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies 76*(3), 1071–1102.

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

List, J. A., A. M. Shaikh, and Y. Xu (2019, December). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*(4), 773–793.

Ma, Y., R. Fairlie, P. Loyalka, and S. Rozelle (2020, April). Isolating the "Tech" from EdTech: Experimental Evidence on Computer Assisted Learning in China. Technical Report w26953, National Bureau of Economic Research, Cambridge, MA.

MHA (2012). 15th Census of India. Place: New Delhi, Delhi Publisher: URL: http://www.censusindia.gov.in/ (last accessed: February 15, 2016). Office of the Registrar General & Census Commissioner, Ministry of Home Affairs, Government of India.

Ministry of Law and Justice (2009, August). Right of Children to Free and Compulsory Education Act, 2009.

Mo, D., Y. Bai, Y. Shi, C. Abbey, L. Zhang, S. Rozelle, and P. Loyalka (2020, September). Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. *Journal of Development Economics 146*, 102487.

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

Mo, D., L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell, and S. Rozelle (2015). Persistence of learning gains from computer assisted learning: Experimental evidence from China. *Journal of Computer Assisted Learning 31*(6), 562–581. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12106.

Mullis, I. V. S. and M. O. Martin (2017). *TIMSS 2019 Assessment Frameworks*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement (IEA).

Muralidharan, K. (2017). Field experiments in education in developing countries. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of economic field experiments*, Volume 2, pp. 323–385. Amsterdam: North Holland.

Muralidharan, K. and A. Singh (2019, June). Improving Schooling Productivity through Computer-Aided Personalization: Experimental Evidence from Rajasthan. Washington, D.C. RISE Annual Conference 2019.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review 109*(4), 1426–60.

NIEPA (2018). *U-DISE Flash Statistics 2016-17*. New Delhi, India: National Institute of Educational Planning and Administration. Google-Books-ID: mUMmAQAAMAAJ.

Pop-Eleches, C. and M. Urquiola (2013, June). Going to a Better School: Effects and Behavioral Responses. *American Economic Review 103*(4), 1289–1324.

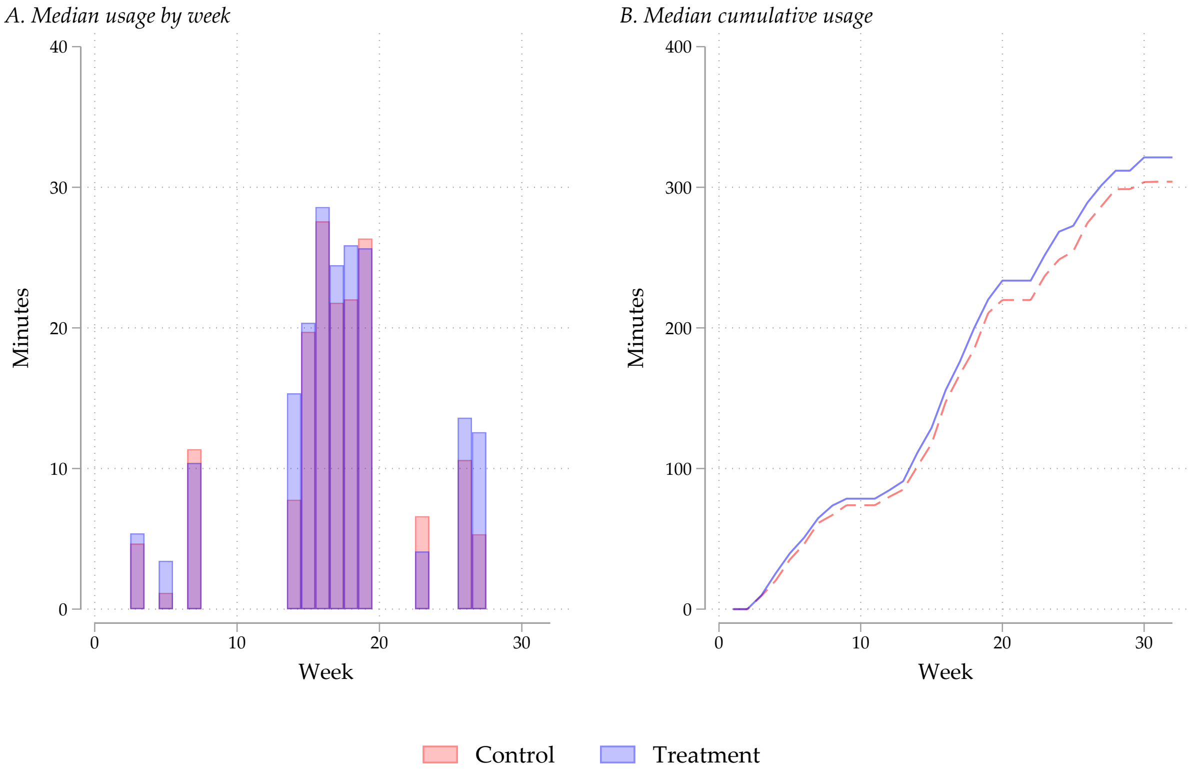
Pritchett, L. and A. Beatty (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development 40*, 276–288.

Rodriguez-Segura, D. (2020, August). Educational Technology in Developing Countries: A Systematic Review. Working Paper 72, University of Virginia, Charlottesville, VA.

Tauson, M. and L. Stannard (2018). *EdTech for learning in emergencies and displaced settings*. London, UK: Save the Children UK.

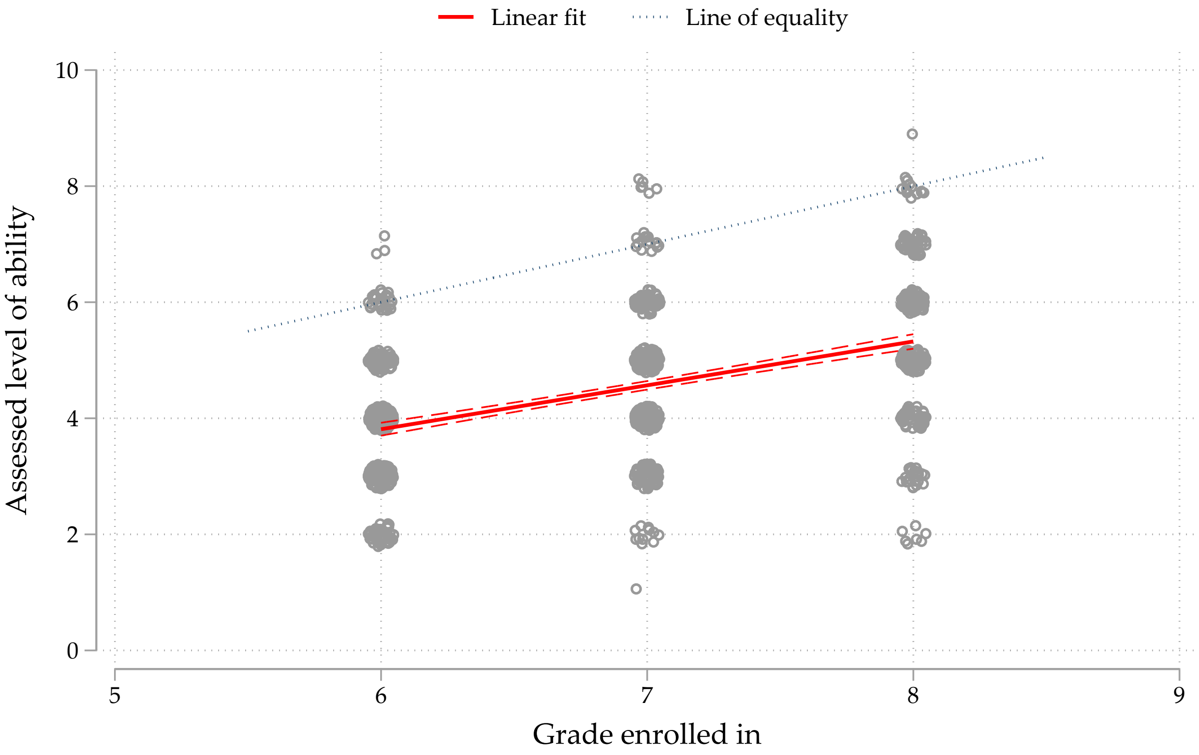
von Hippel, P. T. and L. Bellows (2018, June). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review 64*, 298–312.

Figure 1: Weekly and cumulative time spent on the CAL software during the study



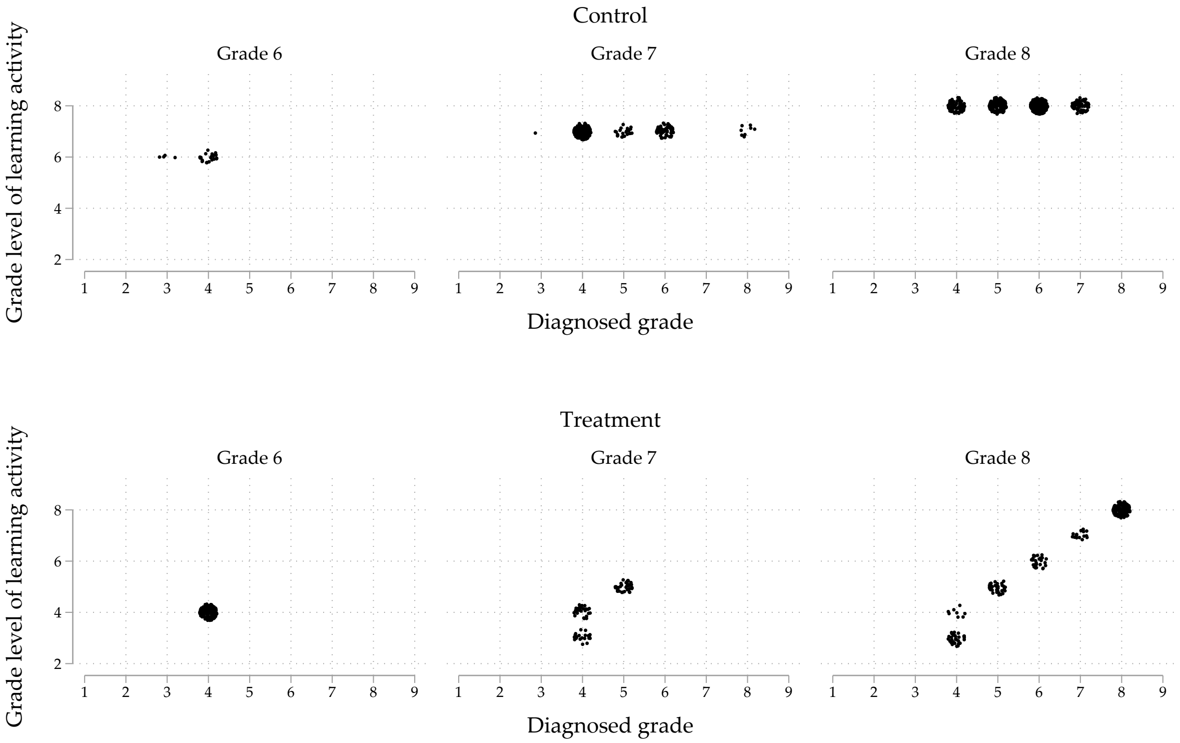
Notes: This figure shows the weekly (panel A) and cumulative (panel B) usage of the CAL platform for the median student, by experimental group. This figure includes all students observed at baseline and endline, regardless of whether they used the software (99.2% of students did). Usage is binned by weeks elapsed since the start of the study (on August 6, 2017).

Figure 2: Students' enrolled grade levels v. their diagnosed grade levels



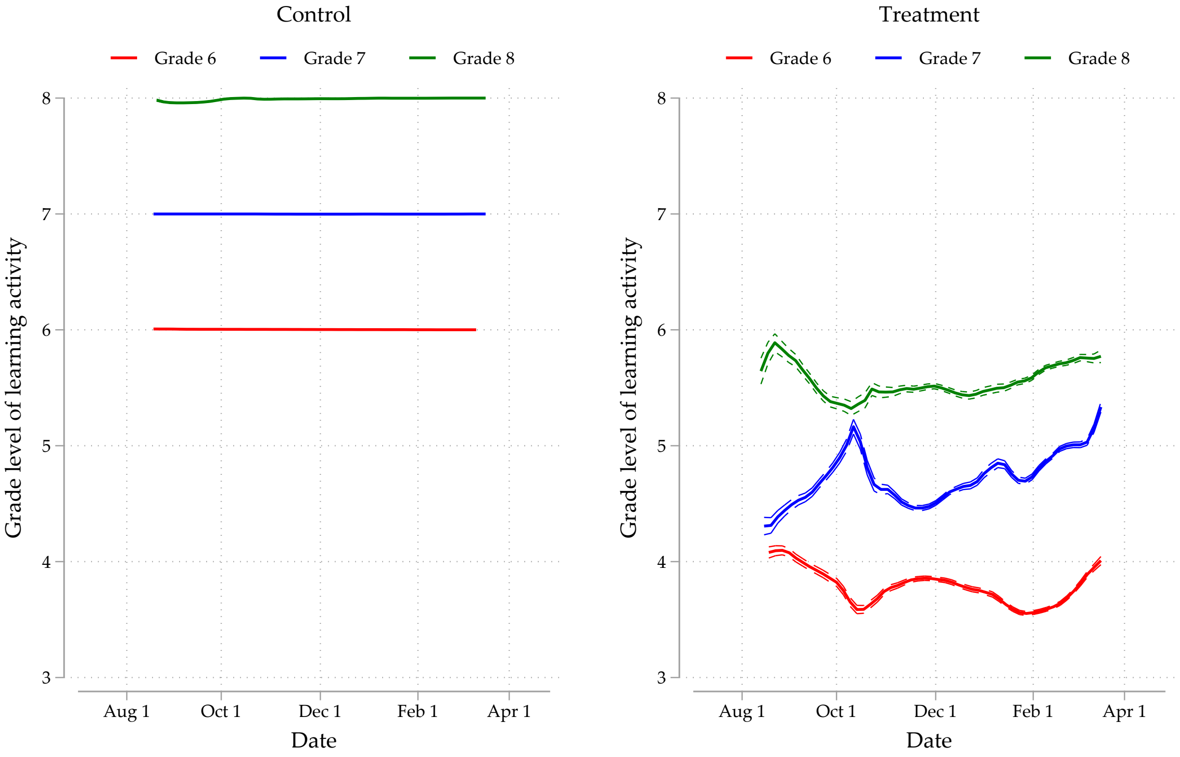
Note: This figure shows the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. These data are from the initial diagnostic test, and do not reflect any instruction provided by Mindspark. We find a general deficit between average attainment and grade-expected norms. We also find a wide dispersion of student achievement, within each grade.

Figure 3: Customization of instruction by CAL software, by treatment status



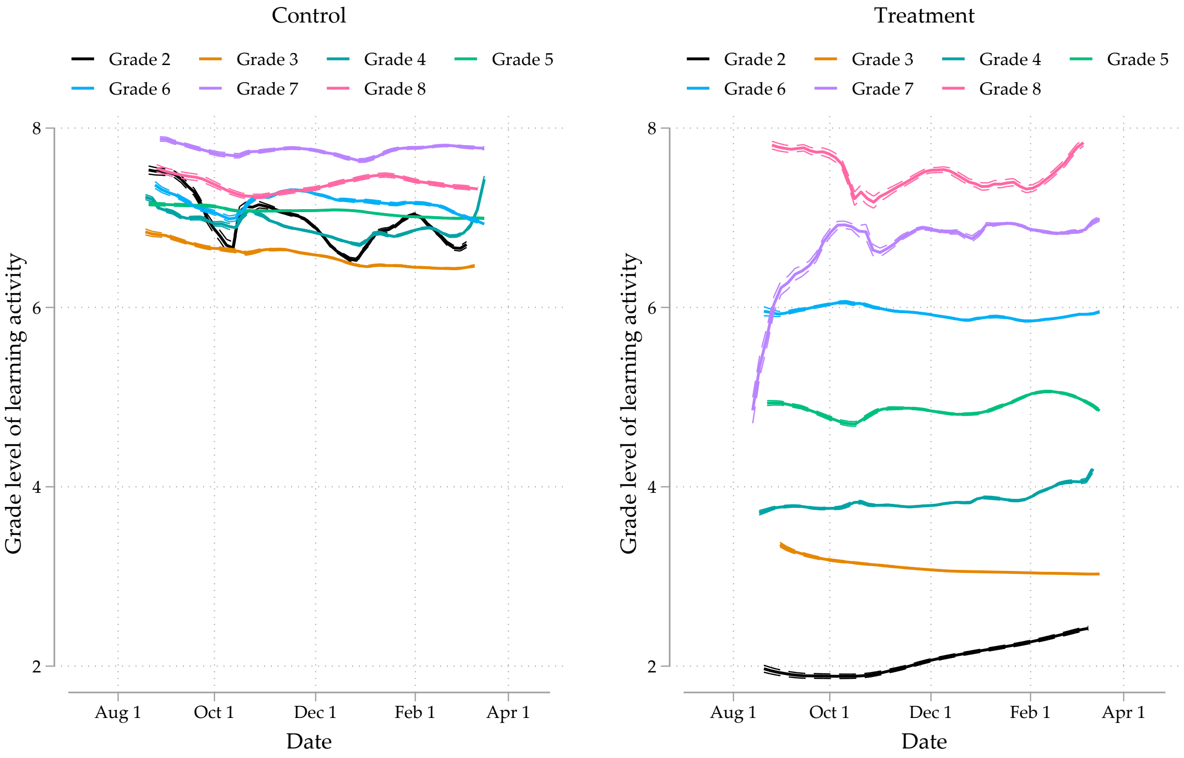
Note: This figure shows, by treatment group, the grade level of learning activities administered by the computer adaptive system to students, on a single day (shortly after activating the study, on August 30, 2017). For simplicity, the figure omits exercises which are also mapped to another, adjacent grade level. In each grade of enrolment, the actual level of student attainment estimated by the CAL software differs widely. In the treatment group, this wide range is covered through the customization of instructional content by the CAL software. In the control group, students only receive materials as per their enrolled grade level.

Figure 4: Dynamic updating and individualization of content, by enrolled grade and experimental group



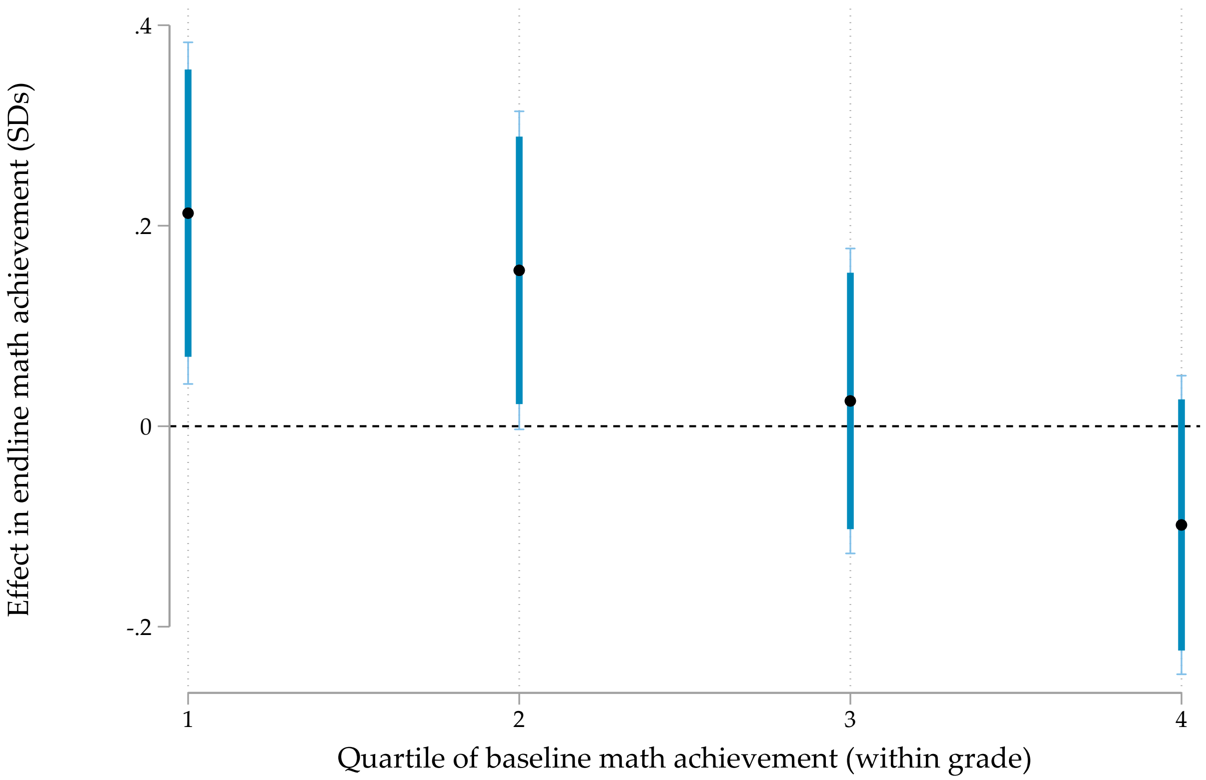
Notes: This figure shows, by treatment group, kernel-weighted local mean smoothed lines relating the level of difficulty of the exercises attempted by students with the date of administration. Separate lines reflect the actual grade of enrolment. The software was activated on August 6, 2017 but its first usage was registered on August 10, 2017. For simplicity, the figure omits learning activities which are also mapped to another, adjacent grade level. Note that 95% confidence intervals are plotted as well but, given the large data at our disposal, estimates are very precise, and the confidence intervals may be too narrow to become visually discernible.

Figure 5: Dynamic updating and individualization of content, by assessed grade and experimental group



Notes: This figure shows, by treatment group, kernel-weighted local mean smoothed lines relating the level of difficulty of the exercises attempted by students with the date of administration. Separate lines reflect the grade level from the software's diagnostic assessment. For simplicity, the figure omits learning activities which are also mapped to another, adjacent grade level. Note that 95% confidence intervals are plotted as well but, given the large data at our disposal, estimates are very precise, and the confidence intervals may be too narrow to become visually discernible.

Figure 6: Heterogeneous ITT effects on math achievement at endline, by quartile of baseline performance



Notes: This figure shows heterogeneity in the intent-to-treat (ITT) effect of personalized learning on students' achievement in math at endline (after 37 weeks), by within-grade quartile of baseline performance. Both panels account for randomization-strata fixed effects. Bars and whiskers show 90-percent and 95-percent confidence intervals, respectively.

Table 1: Balancing checks between experimental groups

	(1) Control	(2) Treatment	(3) Difference
A: Grade-wise distribution (full sample)			
Grade 6	0.34	0.33	
	[0.47]	[0.47]	
Grade 7	0.34	0.34	
	[0.47]	[0.48]	
Grade 8	0.32	0.32	
	[0.47]	[0.47]	
B: Balance tests (full sample)			
Math (IRT-scaled) score	0.02	-0.02	0.05
	[0.99]	[1.01]	(0.04)
Math (percent-correct) score	0.58	0.57	0.01
	[0.17]	[0.17]	(0.01)
Female	0.47	0.49	-0.02
	[0.50]	[0.50]	(0.03)
Attrited from baseline to endline	0.28	0.31	-0.02
	[0.45]	[0.46]	(0.02)
N (students)	762	766	1,528
C: Balance tests (non-attritors)			
Math (IRT-scaled) score	0.08	0.05	0.03
	[0.99]	[1.00]	(0.05)
Math (percent-correct) score	0.59	0.58	0.01
	[0.17]	[0.17]	(0.01)
Female	0.44	0.49	-0.05
	[0.50]	[0.50]	(0.03)
N (students)	547	531	1,078

Notes: This table compares students in the control and treatment experimental groups on their grade-wise enrollment and characteristics: it shows the mean and corresponding standard deviations for each variable (in brackets) and it compares both groups including randomization-strata fixed effects, showing its mean difference and corresponding standard errors (in parentheses). Panel A does not compare enrollment by grade because, due to the stratification strategy, it is comparable across experimental groups by design. Panel B compares students' baseline score and sex (the only two variables collected at baseline) for all students present at baseline. Panel C does the same only for students who were present at baseline and at endline (71% of the total). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: ITT effect of personalization on math achievement at endline

	(1)	(2)	(3)	(4)	(5)
			Math (IRT-scaled) score		
Treatment	0.050	0.053	0.062	0.033	0.056
	(0.054)	(0.055)	(0.040)	(0.045)	(0.038)
Baseline score (std.)			0.72***		0.56***
			(0.024)		(0.029)
Diagnostic score (std.)				0.61***	0.27***
				(0.028)	(0.030)
IPW-adjusted?	No	Yes	No	No	No
N (students)	1,078	1,078	1,078	1,068	1,068
R-squared	0.264	0.277	0.609	0.501	0.639

Notes: This table shows the intent-to-treat (ITT) effect of personalization on students' achievement in math at endline (after 37 weeks). Column 1 shows the simple difference in means; column 2 weights the estimation by each students' inverse probability of participating in the endline; column 3 accounts for students' performance on the independent baseline assessments; column 4 accounts for students' performance on the diagnostic assessments administered by the software upon their first log in; and column 5 accounts for students' baseline performance on both assessments. All estimations include randomization-strata fixed effects. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: ITT effect of personalization on math achievement at endline, by content and cognitive domain

	A. All students					
	(1)	(2)	(3)	(4)	(5)	(6)
	Numbers	Geometry	Data	Knowing	Applying	Reasoning
Treatment	0.009	0.014	0.023*	0.009	0.018**	0.012
	(0.008)	(0.010)	(0.012)	(0.008)	(0.008)	(0.014)
Baseline score	0.122***	0.133***	0.134***	0.118***	0.124***	0.168***
	(0.005)	(0.006)	(0.007)	(0.005)	(0.005)	(0.009)
FWER-adj. p-value	0.592	0.43	0.239	0.51	0.182	0.38
N (students)	1078	1078	1078	1078	1078	1078
R-squared	0.503	0.465	0.411	0.471	0.534	0.418

	B. Low-performing students					
	(1)	(2)	(3)	(4)	(5)	(6)
	Numbers	Geometry	Data	Knowing	Applying	Reasoning
Treatment	0.048***	0.025	0.046*	0.032*	0.051***	0.030
	(0.018)	(0.022)	(0.026)	(0.019)	(0.018)	(0.032)
Baseline score	0.127***	0.136***	0.119***	0.121***	0.133***	0.130***
	(0.013)	(0.016)	(0.019)	(0.013)	(0.013)	(0.023)
FWER-adj. p-value	0.298	0.85	0.459	0.744	0.068	0.858
N (students)	1078	1078	1078	1078	1078	1078
R-squared	0.515	0.476	0.422	0.480	0.541	0.424

Notes: This table shows the intent-to-treat (ITT) effect of personalization on students' achievement in each content (columns 1-3) and cognitive (columns 4-6) domain at endline (after 37 weeks). All estimations include randomization-strata fixed effects. Panel A provides average ITT effects among all students. Panel B uses interactions (not shown) to report ITT effects among students in a grade-level's bottom quartile, as per students' performance on the baseline assessment. The last row of each panel shows p-values for the treatment coefficient, adjusted for multiple hypothesis testing that asymptotically controls the family-wise error rate (FWER), following List et al. (2019). Adjustments account for treatment effects in all quartiles, including in those not reported on in the table (i.e., for 24 tests). * significant at 10%; ** significant at 5%; *** significant at 1%.

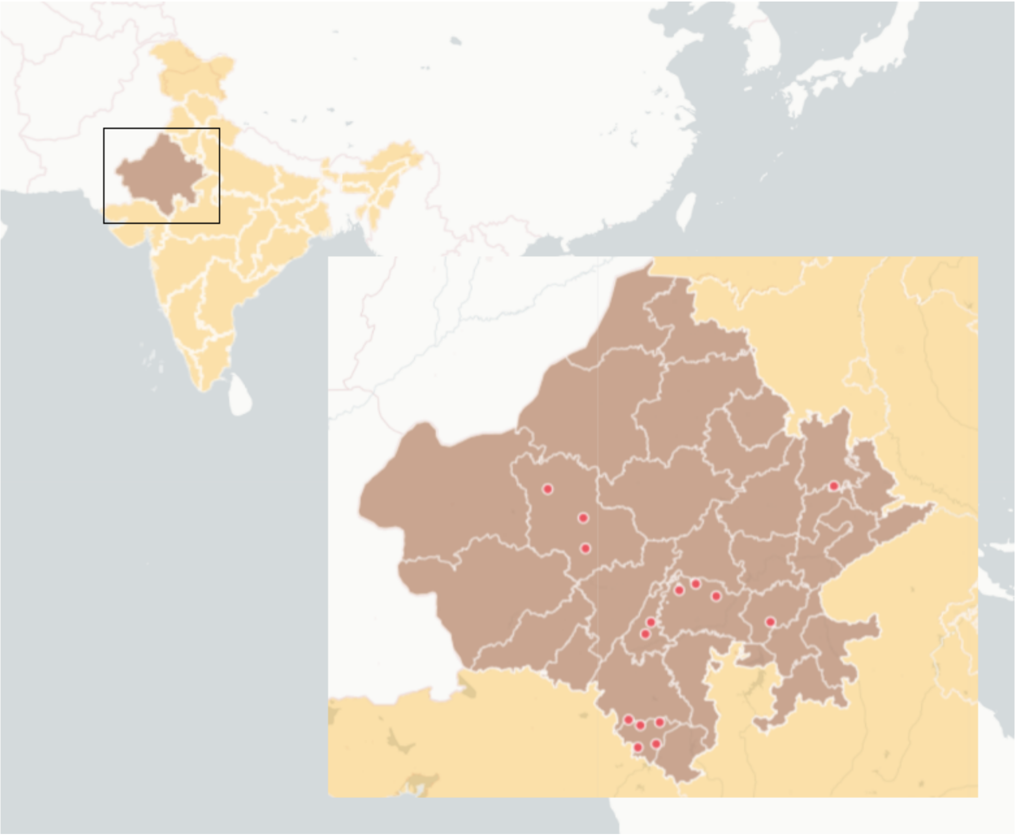
Table 4: Heterogeneous ITT effects on math achievement at endline, by students' baseline performance

	(1)	(2)	(3)	(4)
	Math (IRT-scaled) score			
Treatment	0.278***	0.215**	0.229**	0.192*
	(0.085)	(0.088)	(0.097)	(0.112)
	[0.026]	[0.081]	[0.12]	[0.344]
Baseline (percentile)	0.024***	0.032***		
	(0.001)	(0.003)		
Treatment X Baseline	-0.004***			
	(0.001)			
	[0.062]			
Quartile 2		-0.287**		
		(0.111)		
Quartile 3		-0.514***		
		(0.169)		
Quartile 4		-0.638***		
		(0.235)		
Treatment X Quartile 2		-0.057		
		(0.122)		
		[0.943]		
Treatment X Quartile 3		-0.182		
		(0.119)		
		[0.548]		
Treatment X Quartile 4		-0.338***		
		(0.118)		
		[0.076]		
Diagnostic (percentile)			0.021***	0.014***
			(0.001)	(0.002)
Treatment X Diagnostic			-0.004**	
			(0.002)	
			[0.184]	
Student is 2-3 levels behind				0.359***
				(0.109)
Student is 0-1 levels behind				0.685***
				(0.160)
Treatment X 2-3 levels behind				-0.167
				(0.128)
				[0.638]
Treatment X 0-1 levels behind				-0.268*
				(0.150)
				[0.41]
N (students)	1,078	1,078	1,068	1,068
R-squared	0.599	0.606	0.491	0.498

Notes: This table shows the intent-to-treat (ITT) effect of personalization on students' achievement in math at endline (after 37 weeks) by baseline performance on the study's independent tests, and on the software's diagnostic test. Baseline performance is expressed within grade-levels, as percentiles (column 1) and as quartile indicator variables (column 2). Performance on the diagnostic test is expressed within grade-levels, as percentiles (column 3) and as indicator variables for the number of grade-levels students lagged behind (column 4). All estimations include randomization-strata fixed effects. * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors in parentheses; p-values in brackets, adjusted for multiple hypothesis testing that asymptotically controls the familywise error rate (FWER), following List et al. (2019). Adjustments conservatively account for *all* (prespecified) tests of heterogeneous effects, including those documented in Table A.4 (i.e., for 16 tests).

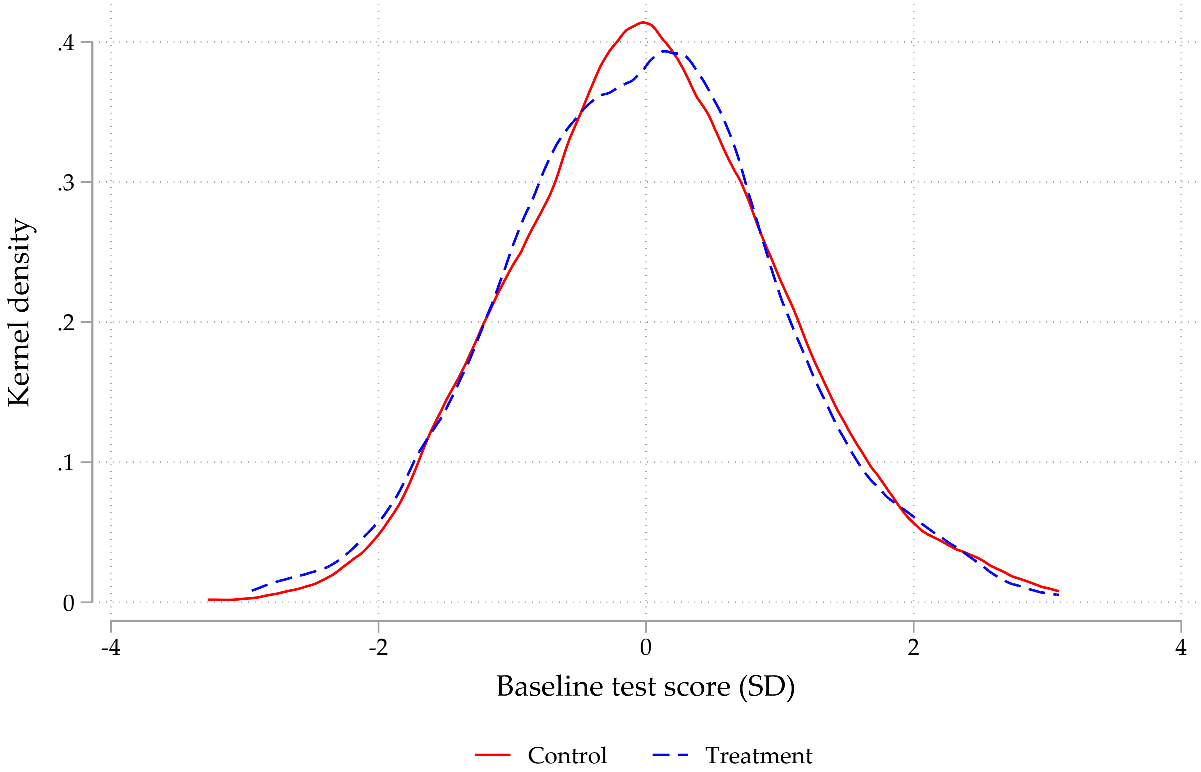
Appendix A Additional graphs and tables

Figure A.1: Map of study districts and schools



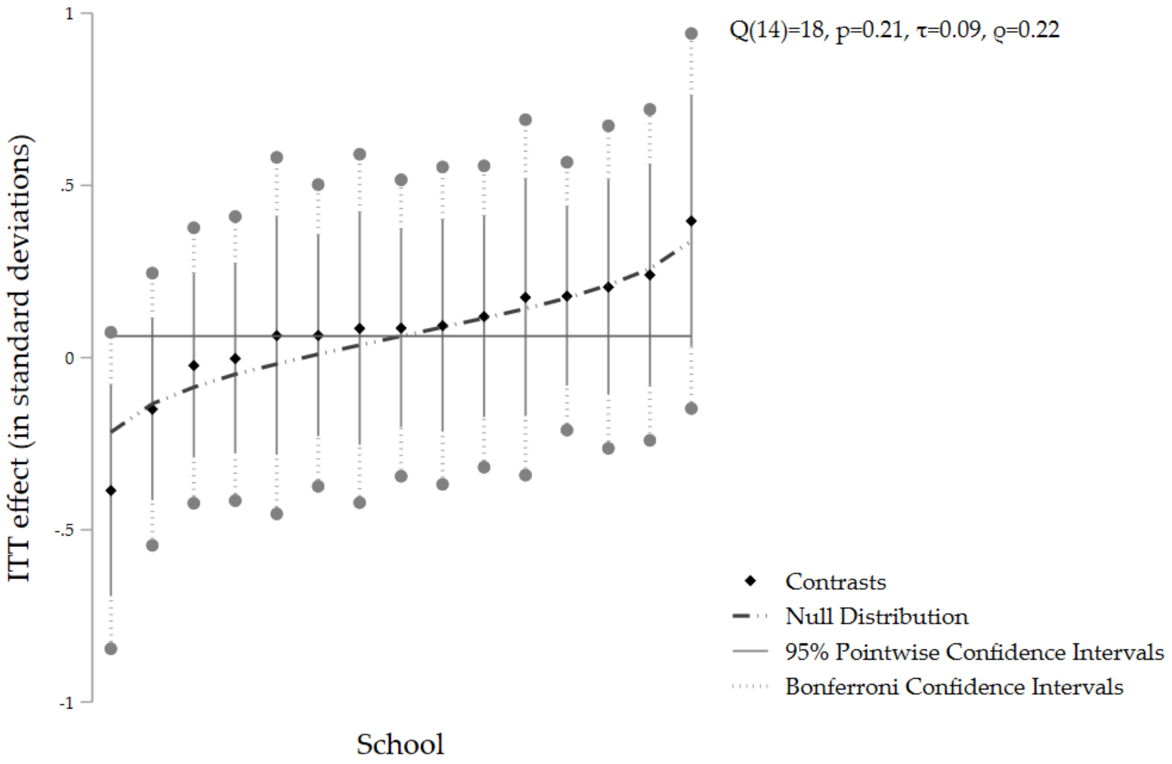
Note: This figure shows the state of Rajasthan (in brown), and the location of study schools (in red).

Figure A.2: Distribution of math (IRT-scaled) scores by experimental group at baseline



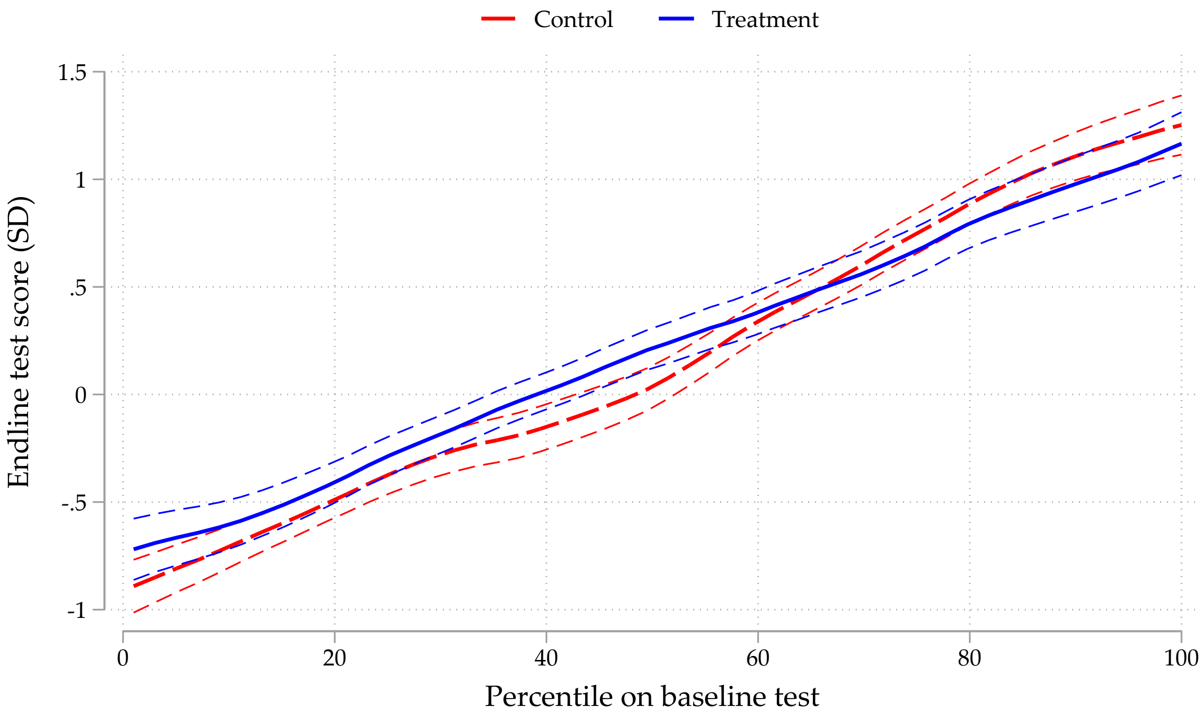
Notes: This figure shows the distribution of scores in the baseline assessment of math for control and treatment students. Scores were scaled using a two-parameter logistic Item Response Theory (IRT) model. This figure includes all students present at baseline and at endline.

Figure A.3: Heterogeneous ITT effects on math achievement at endline, by school



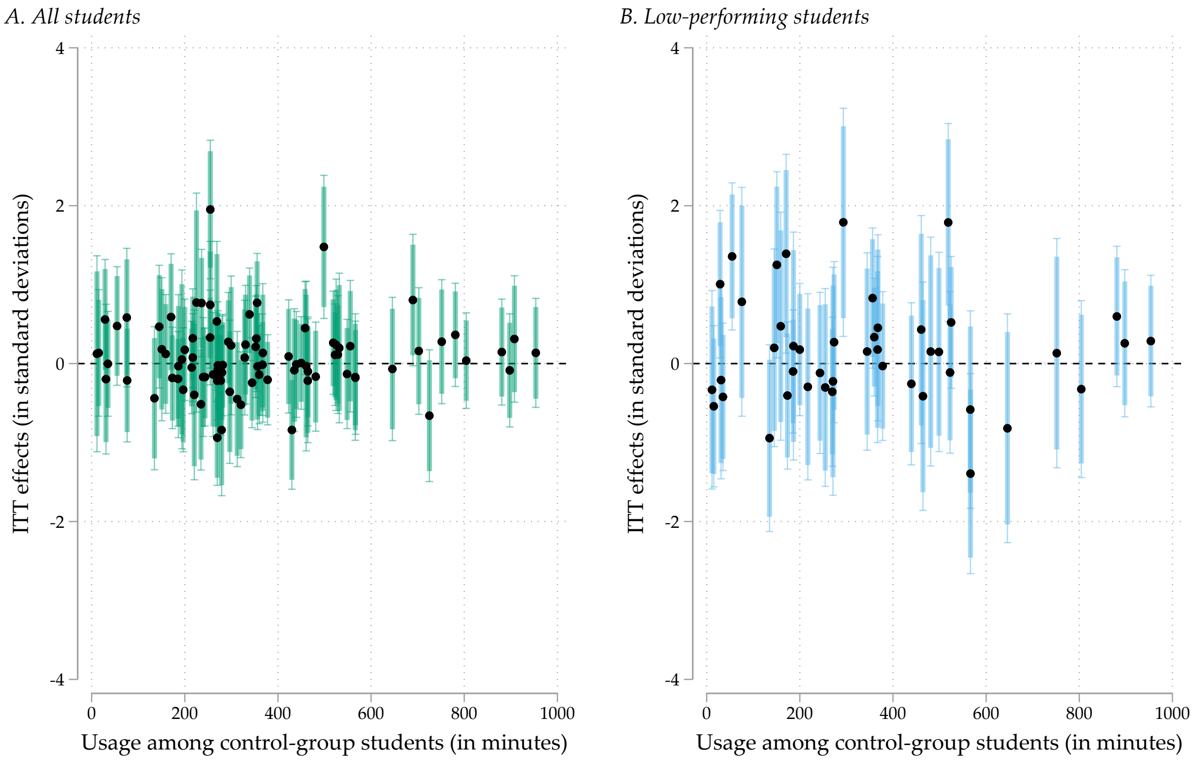
Notes: This figure provides a "caterpillar plot" of ITT effects by school (cf. von Hippel and Bellows 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution of "effects" that can be expected due to error. τ is the heterogeneity standard deviation. Q refers to Cochran's Q statistic, which follows a χ^2 distribution, and p reports on the corresponding p-value for a test of the null hypothesis of no heterogeneity. ρ estimates the reliability; that is, the share of variance in estimates that is attributable to heterogeneity (rather than error). The estimation controls for student baseline achievement and randomization-strata fixed effects.

Figure A.4: Non-parametric investigation of treatment effects by (within-grade) percentiles on the baseline test



Note: The figure presents kernel-weighted local mean smoothed plots which relate endline test scores to within grade-level percentiles in the baseline achievement, separately for the treatment and control groups, alongside 95% confidence intervals. In approx. the bottom two quartiles of baseline achievement, treatment group students score higher in the endline test than the control group; there are no discernable differences for the top half of the distribution.

Figure A.5: Dose-response relationship



Notes: This figure shows heterogeneity in the intent-to-treat (ITT) effect of personalized learning exercises on students' achievement in math at endline (after six months) by randomization stratum, for all students (panel A) and students in the bottom quartile of baseline achievement within their grade level (panel B). Bars and whiskers show 90-percent and 95-percent confidence intervals, respectively.

Table A.1: Number and percentage of attempted exercises by content domain and topic

Topic	(1) Number of exercises	(2) Percentage of exercises (total)
Panel A. Numbers	*39,023*	*95.01%*
Whole Number Concepts	11,414	27.79%
Whole Number Operations	7,964	19.39%
Real Numbers	6,126	14.91%
Integers	3,700	9.01%
Number theory	3,322	8.09%
Basic Algebra	2,977	7.25%
Fractions	1,769	4.31%
Decimals	1,725	4.20%
Ratio and Proportion	15	0.04%
Percentages and commercial maths	6	0.01%
Exponents	5	0.01%
Panel B. Geometry	*1,863*	*4.53%*
Measurement	1,021	2.49%
Geometry	732	1.78%
Area	104	0.25%
Volume and Surface Area	6	0.01%
Panel C. Data	*188*	*0.46%*
Probability and Data Analysis	188	0.46%

Notes: This table shows the number of exercises that study participants across both experimental groups attempted on the CAL software, as well as the percentage of the total that the number represents. Panel A shows topics related to numbers, panel B shows topics related to geometry, and panel C shows topics related to data.

Table A.2: Lee bounds estimates of ITT effect of personalization on math achievement at endline

	(1) Math (IRT-scaled) score
Lower	0.023
	(0.077)
Upper	0.104
	(0.077)
Lower 95% CI	-0.107
Upper 95% CI	0.238

Note: This table shows the Lee (2009) bounds on the intent-to-treat (ITT) effect of personalization on students' achievement in math at endline (after 37 weeks). As the dependent variable, we use residuals from a regression of endline test scores on baseline test scores and randomization fixed effects, to keep our analysis of bounds analogous to the main ITT effects. The bounds are tightened within school-by-grade cells. Analytic standard errors are shown in parentheses.

Table A.3: ITT effect of personalization on math achievement at endline, by repeated and non-repeated items

	All students		*Low-performing students*	
	(1) Repeated items (proportion-correct) score	(2) Non-repeated items (proportion-correct) score	(3) Repeated items (proportion-correct) score	(4) Non-repeated items (proportion-correct) score
Treatment	0.008	0.025***	0.028*	0.068***
	(0.007)	(0.009)	(0.016)	(0.021)
Baseline score	0.125***	0.134***	0.129***	0.126***
	(0.005)	(0.006)	(0.012)	(0.015)
N (students)	1,078	1,078	1,078	1,078
R-squared	0.571	0.495	0.576	0.505

Notes: This table shows the intent-to-treat (ITT) effect of personalization on students' achievement in items administered in both baseline and endline (which we call "repeated items" in columns 1 and 3) and items that were first introduced in the endline (which we call "non-repeated items" in columns 2 and 4) after 37 weeks. All estimations include randomization-strata fixed effects. Columns 1 and 2 provide average ITT effects among all students. Columns 3 and 4 use interactions (not shown) to report ITT effects among students in a grade-level's bottom quartile, as per students' performance on the baseline assessment. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Heterogeneous ITT effects of personalization on math achievement at endline, by students' sex and enrolled grade

	(1)	(2)
	Math (IRT-scaled) score	
Treatment	0.054	0.017
	(0.055)	(0.065)
	[0.88]	[0.937]
Baseline score (std.)	0.72***	0.72***
	(0.025)	(0.024)
Student is female	-0.031	
	(0.058)	
Treatment X Female	0.021	
	(0.082)	
	[0.81]	
Treatment X Grade 7		0.078
		(0.093)
		[0.918]
Treatment X Grade 8		0.064
		(0.10)
		[0.936]
N (students)	1,078	1,078
R-squared	0.609	0.609

Notes: This table shows the intent-to-treat (ITT) effect of personalization on students' achievement in math at endline (after 37 weeks) for female students (column 1) and students enrolled in different grades (column 2). All estimations include baseline achievement and randomization-strata (i.e., grade) fixed effects (coefficients not shown). Standard errors in parentheses; p-values in brackets, adjusted for multiple hypothesis testing that asymptotically controls the family-wise error rate (FWER), following List et al. (2019). Adjustments conservatively account for all (prespecified) tests of heterogeneous effects, including those documented in Table 4 (i.e., for 16 tests). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: ITT effect of personalization on usage of CAL platform

	Number of sessions completed (log)	Total minutes spent on CAL platform (log)	
	(1)	(2)	(3)
Treatment	0.008	0.028	0.023
	(0.026)	(0.025)	(0.017)
Baseline score	0.062***	0.055***	0.012
	(0.016)	(0.015)	(0.011)
Number of sessions completed (log)			0.695***
			(0.021)
N (students)	1,069	1,069	1,069
R-squared	0.695	0.798	0.905

Notes: This table shows the intent-to-treat (ITT) effect of personalization on the (natural logarithm of) number of sessions that students completed (column 1), on the (natural logarithm of) minutes they spent on the CAL platform (column 2), and on that same number holding the number of sessions completed constant (column 3). All estimations include randomization-strata fixed effects. The estimations excludes nine students who did not spend any time on the software. * significant at 10%; ** significant at 5%; *** significant at 1%.