Which Students Benefit from Computer-Based Individualized Instruction? Experimental Evidence from a Math Software in Public Schools in India*

Andreas de Barros†
Massachusetts Institute of Technology

Alejandro J. Ganimian‡
New York University

Abstract

This is one of the first studies to evaluate the impact of computer-based individualized instruction in a developing country. We randomly assigned 1,528 students in grades 6-8 in 15 "model" public schools in Rajasthan, India who were using a computer-adaptive learning software to: a control group, in which they were only able to access the activities for their enrolled grade level; or a treatment group, in which they were able to access exercises appropriate for their performance level. After nine months, computer-based individualized instruction had a null average effect on math achievement. However, treatment students with low initial performance outperformed their control counterparts by 0.22 standard deviations. Our results suggest that computer-based individualized instruction is most beneficial for low-performing students.

JEL codes: C93, I21, I22, I25

Keywords: computer-aided learning, India, computer-based individualized instruction, math instruction

---

* 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. E-mail: debarros@mit.edu.

‡ Assistant Professor of Applied Psychology and Economics, New York University Steinhardt School of Culture, Education, and Human Development. E-mail: 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 considerable variation in students' preparation for school within each grade (see, for example, Andrabi et al., 2007; Das & Zajonc, 2010; Duflo et al., 2011; Pritchett & Beatty, 2015; Muralidharan et al., 2019). This pattern is expected to be exacerbated by the recent school closures due to the ongoing pandemic (Azevedo et al., 2020; Kaffenberger, 2020; Kaffenberger & Pritchett, 2020; Angrist et al., 2021).

School systems have sought to address heterogeneity in student preparation for schooling in two main ways: by asking teachers to provide differentiated instruction (i.e., dividing students into groups based on their performance within the classroom and assigning activities that cater to each group) or computer-adaptive learning (i.e., providing students with access to a software that dynamically adjusts to their level and rate of learning). Individualized instruction—known in some circles as "teaching at the right level"—has improved student learning when it is implemented as intended, but teachers have often 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., 2007; Banerjee et al., 2010; Banerjee et al., 2017). Computer-adaptive learning (CAL) has also yielded promising results, but the multiple components that CAL software products typically combine make it challenging to understand the role that individualized instruction plays in its effectiveness (see, e.g., Banerjee et al., 2007; Muralidharan et al., 2019; Muralidharan & Singh, 2020).

To advance existing knowledge on how to address heterogeneity in students' preparation for learning, we conducted a randomized evaluation to investigate whether individualized

instruction, when delivered through a CAL software, improves achievement. Specifically, in the present paper, we present one of the first experimental estimates of the effect of computer-based individualized instruction on math achievement in a developing country, using a software product (called "Mindspark") that has been previously found to improve scores on standardized tests both when delivered before/after school and during school hours (Muralidharan et al., 2019; Muralidharan & Singh, 2020). Unlike these prior evaluations, which have estimated the effect of the Mindspark software as a whole—including, but not limited to the individualized instruction feature—ours isolates this feature. It focuses on whether the fact that students who are diagnosed to lag behind curricular expectations are presented with material that they were supposed to learn in earlier grades improves their performance on math tests. We randomly assigned students in grades 6 to 8 in public "model" schools (which serve disadvantaged areas but select students based on entrance exams) 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; or (b) a treatment group, in which they were assigned exercises appropriate for their individual preparation level.

We report three main sets of results. First, as it has been shown in other settings, the learning levels of most students were several levels behind grade-level expectations (based on a diagnostic test that all students were required to take when they first logged into the software). Thus, the students in our context, much like those in many low- and middle-income countries, stood to benefit considerably from individualized instruction.

Second, computer-based individualized instruction 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, but this difference was not statistically significant. In fact, based on the 95% confidence interval, we could rule out effects below -0.02 SDs and above 0.13 SDs. Further, we find that effects remain close to zero even at very high levels of exposure to the intervention. To put these effects in context, the first efficacy trial of the CAL software that we used to test the effect of computer-based individualized instruction found that the entire program (including, but not limited to the individualized instruction component) improved math test scores by 0.59 SDs in math and 0.36 SDs in Hindi after only 4.5 months (Muralidharan et al., 2019). If individualized instruction was indeed driving these gains, it seems reasonable to expect effects well above 0.13 SDs after 9 months of interaction with the software, even considering the differences in sampling and dosage.

Third, computer-based individualized instruction had a positive effect for low performers (i.e., those in the bottom quartile of the within-grade baseline math achievement distribution) in the treatment group, who outperformed their control counterparts in the math assessment that we developed by 0.22 SDs. This effect remains statistically significant at the 10% level after we account for multiple hypothesis testing. We also estimate the effect of the intervention on their percentile rank in the diagnostic test that students complete when they first log into the CAL platform and find further evidence of an interaction effect between the intervention and baseline achievement. These effects are not statistically significant once we account for multiple hypothesis testing, but they are consistent with the effects we observe on the test that we designed.

The rest of the paper is structured as follows. Section 1 reviews the literature on CAL in general and technology-enabled individualized instruction in particular. Section 2 presents the context, study design, and intervention. Section 3 describes the data. Section 4 discusses the

empirical strategy. Section 5 reports the results. Section 6 discusses implications for research and policy.

1. Prior Research

Prior experimental and quasi-experimental evaluations of education technology interventions in developed and developed countries indicate that, while providing students with free hardware (e.g., laptops) has generally had null—and in some cases, *negative*—effects on their achievement (see, e.g., Angrist & Lavy, 2002; Leuven et al., 2007; Barrera-Osorio & Linden, 2009; Malamud & Pop-Eleches, 2011; Fairlie & Robinson, 2013; Beuermann et al., 2015; Cristia et al., 2017), offering them opportunities to interact with educational software—often, as a complement to teacher-led instruction—has typically had more encouraging results. Specifically, software that focuses on getting students to review the material taught by their teacher on a given week has traditionally produced small-to-moderate improvements in test scores and/or school grades (see, e.g., Lai et al., 2012; Lai et al., 2013; Mo et al., 2014; Lai et al., 2015; Mo et al., 2015; Mo et al., 2016), whereas software that provides some degree of differentiation or individualized instruction has had moderate-to-large achievement gains (see Banerjee et al., 2007; Muralidharan et al., 2019; Muralidharan & Singh, 2020). This research suggests that technological innovations that purposefully address a binding constraint to student learning (e.g., heterogeneity in students' preparation for schooling) have much greater chances of success.

Evaluations of software products, however, cannot causally identify the effect of any individual feature. Each product that includes a individualized instruction feature also includes many other features, such as high-quality content vetted by pedagogical experts, feedback on incorrect answers, opportunities for independent practice, and both "game-based" (e.g.,

labyrinths) and “gamified” (e.g., leaderboards) elements. Each feature may impact achievement positively (e.g., independent practice may reinforce students’ procedural knowledge; see de Barros et al., 2022) or negatively (e.g., gamification may increas students' anxiety; see Araya et al., 2019) and interact with another feature in ways that reinforce or offset its effects. An evaluation of a software product captures the aggregation of all those main and interaction effects.

To our knowledge, our study is the first to experimentally evaluate the effect of technology-enabled individualized instruction in a developing country (i.e., to randomly assign students to a version of the software in which this feature is included or to another one in which this feature is excluded). Only one other study adopted a similar design, but it was conducted in a developed country (see Van Klaveren et al., 2017). Experimentally evaluating the impact of individualized instruction in low- and middle-income countries is particularly important because students in those settings are far more likely than their counterparts in high-income nations to lag behind curricular standards and vary in their achievement (see, e.g., MIA, 2015; Uwezo, 2019; ASER Pakistan, 2020; ASER, 2021). Thus, computer-based individualized instruction addresses a constraint to learning that is more binding in the developing world.

In theory, computer-based individualized instruction should benefit all students equally. If a software adjusts the difficulty of the material presented to each student *dynamically* (i.e., as a function of that student’s performance on each set of activities), it could improve his/her achievement by presenting him/her with activities that build on his/her understanding and remedy his/her gaps. Indeed, this is what prior evaluations of software products that include individualized instruction suggest. Both Banerjee et al. (2007) and Muralidharan et al. (2019)

find that two products with different levels of differentiation had comparable effects for low- and high-performing students in India.

Yet, existing evidence also indicates that the quality of counterfactual instruction matters. Muralidharan et al. (2019) note that, while the software that they evaluated in Delhi, India yielded similar *absolute* gains for students along the achievement distribution, students in the bottom third of the distribution reaped higher *relative* gains than those in the top third, since control students in this group did not improve at all during the 4.5 months of the evaluation. Likewise, Linden (2008) found that the same software that had produced moderate-to-large gains when evaluated by Banerjee et al. (2007) in public schools in Gujarat, India had null effects when deployed in a well-functioning network of schools run by a non-profit in the same state. This pattern of results suggests that individualized instruction has a greater margin for impact for low-performing students when teacher-led instruction fails to cater to the needs of these students. In this study, we investigate such individualization in the context of computer-adaptive learning, and devote particular attention to those students who lag farthest behind.

2. Experiment

Context

Schooling in India is compulsory and free from ages 6 to 14 (MLJ, 2009). Primary education runs from grades 1 to 5 and upper primary runs from grades 6 to 8. In the 2016-2017 school year, the Indian school system included 840,241 primary schools, 287,265 upper-primary schools, and 48,543 primary schools with secondary grades (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 individualized instruction (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 subtract a two-digit number from another two-digit number and just 29% could divide a three-digit number by a two-digit number, compared to 24% and 35% of sixth-graders in all of rural India (ASER, 2018).

Specifically, we conducted our study in "model" public schools, which 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 disadvantaged areas of the state (as opposed to the entire state), they only cover grades 6 to 12 (instead of grades 1 to 12), their medium of instruction is English (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 pass an entrance exam to gain admission (Kumar, 2020). 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 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 weekly slots for computer-science lessons) that make it easier to deploy an educational software.

Yet, the same characteristics that make model schools propitious for our intervention also limit the generalizability of our findings. Due to their focus on students who have completed primary school, who speak and write in English, and who can pass an entrance exam, these schools likely serve students with not only higher average achievement, but also less variability in their performance than traditional public schools. Therefore, the potential contribution of computer-based individualized instruction for the *average student* in these schools is likely to be lower, given that individualized instruction seeks to address heterogeneity in students' preparation. Further, due in part to their curriculum and infrastructure standards, these schools may offer more opportunities for students to learn than traditional public schools. Thus, the potential for technology-enabled individualized instruction to improve student learning is likely more limited, given that it seeks to compensate for the absence of differentiated-instruction strategies by teachers. Put differently, the average impact of technology-enabled individualized instruction may be higher in traditional public schools.

Sample

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

Rajsamand, and Udaipur (see Figure A.1 in Appendix A). We selected schools based on three criteria: (a) they had to have fully constructed buildings; (b) they had to have space for a computer lab; and (c) they had to have running electricity. In theory, all 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. 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 found 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 between groups is not statistically significant. To verify that our pattern of results is not affected by attrition, we included both inverse-probability weighted (IPW) estimates and Lee (2009) bounds.

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).[1] We describe the differences across groups in the Intervention section. To maximize comparability across experimental groups, we randomly assigned participants to experimental groups at the student level (instead of at the school or classroom levels) and we stratified the randomization within each school-by-grade-by-section combination (e.g., one lottery included students in school 1, grade 6, and section "A"). Principals, teachers, and students were "blind" to each student's experimental condition.

It is possible, but highly unlikely, that under this randomization strategy teachers notice that treatment students are improving because of this intervention and engage in compensatory behavior by reinforcing their instruction of control students. First, for this to happen teachers would have to know how the learning levels of their students evolved during the school year. Yet, teachers in India misestimate the skills of students in their classroom by a wide margin: according to a recent study in the state of Maharashtra, 84% of teachers in grades 5 and 6 overestimated the performance of their students in an independent math assessment; in fact, the average teacher misestimated his/her students' score by 24 percentage points or 126% of the within-class standard deviation in math achievement (Djaker et al., 2022). Second, for this to happen teachers would have to adopt different instructional strategies for low performers. Teachers in India, however, are widely known to rarely cater to the needs of their students. Classroom observations have consistently shown that these teachers use the same materials for all students and spend most of their lessons using whole-classroom approaches (Bhattacharjea et al., 2011; Sankar & Linden, 2014; Ganimian et al., 2022). This is precisely what makes differentiated or individualized instruction necessary in this context (Ganimian & Djaker, 2022). Third, for this to happen, teachers would need to identify the treatment and control students. Yet, teachers are blind to experimental groups and all students are interacting with the software.

The randomization of students within the same classroom maximizes statistical power, but its main drawback is that it allows for spillovers across students in the same classroom. In theory, if treatment students (who had access to the intervention) work together with control students (who did not have access to the intervention) on math exercises, treatment-control comparisons could under-estimate the effect of the intervention on math achievement of the former (if they transferred some of their knowledge to the latter).

We believe such spillovers are possible but unlikely to be a major concern. The individualized instruction feature in the educational software benefits each student by presenting him/her with material that addresses gaps in his/her knowledge and the underlying misconceptions (see Appendix D of Muralidharan et al., 2019). Therefore, for spillovers to offset differences between treatment and control students, three conditions would have to be met. First, treatment and control students would have to regularly work together—something that rarely occurs in classroom observations in India (see, e.g., Bhattacharjea et al., 2011; Sankar & Linden, 2014; Sinha et al., 2016; World Bank, 2016a). Second, the treatment student and the control student would have to have similar gaps in knowledge and/or misconceptions—something that is contradicted by the wide dispersion of grade levels of the activities presented by the platform on any given day (which we display in Figure 3 and discuss in the Results section). Third, treatment students would have to be (approximately) as effective in addressing their control peers' misconceptions as the individualized instruction feature of the software is with treatment students—something that is belied not only by the expertise and 10 years of iteration that has gone into the design of the software, but also by the evidence on peer-to-peer learning in developing countries (see, e.g., Beuermann et al., 2013; Berlinski & Busso, 2017).

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 similar across experimental groups (see Figure A.2).

Intervention

We provided all students in our study with a CAL software called "Mindspark", which focused on math instruction. The software can also provide language instruction, but this

function was deactivated in our study. It 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 is also achieved at lower costs (Muralidharan & Singh, 2020).

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 individualized instruction 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 adjust to each student's 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 focus on the content prescribed for their enrolled grade level (see, e.g., Carrillo et al., 2011; Lai et al., 2012; Lai et al., 2013; Mo et al., 2014; Lai et al., 2015; Mo et al., 2015; Mo et al., 2020). In fact, this version of the software also resembled business-as-usual teacher-led instruction in India, where classroom observations have found that teachers use the same materials and instructional approaches for all students, regardless of their preparation level (Bhattacharjea et al., 2011; Sankar & Linden, 2014; Sinha et al., 2016; World Bank, 2016b). This version allows for some degree of individualized instruction within grade-appropriate materials, but students are not presented with material for lower grades regardless of how far behind they lag in their performance. For example, if a control group student gets introduced to fractions and struggles, he/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 learning fractions). This feature of our study allows us to shed light on the potential contribution of the individualized instruction feature to computer-aided learning.

A few of the educational software products that have been evaluated in developing countries include some degree of individualized instruction (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 individualized instruction of both the content and difficulty of the material.

3. 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).

Student achievement

We administered student assessments of math at baseline (before the intervention) and endline (approximately nine months after the start of the intervention). 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 Mathematics and Science Study (TIMSS) for grade 4 (IEA, 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

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), including items from a wide array of grade levels (i.e., grades 3 to grade 8) to make sure that we could capture impacts on learning outcomes on foundational skills. We created three versions of the assessment at baseline and four versions at endline to prevent students from cheating.[2]

To map both the baseline and endline assessments onto the same scale, we used a non-equivalent anchor test (NEAT) design (see Kolen & Brennan, 2004). We included 11 items in common across both rounds of assessment (known as an "anchor test") and then we fit a two-parameter logistic Item Response Theory (IRT) model, which accounts for differences in the difficulty and "discrimination" (capacity to distinguish between otherwise similarly performing examinees) across items (Yen & Fitzpatrick, 2006), pooling data from both assessment rounds. This process places the baseline and endline results onto a common scale.

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.

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 the CAL platform, where we can link students to sessions using their unique login credentials); (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.

4. Empirical strategy

We estimate the effect of the offer of computer-based individualized instruction (i.e., the "intent-to-treat" or ITT effect) by fitting the following model:

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

where Y_{igs}^{t} 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 $\alpha_{r(gs)}$ is a stratum fixed effect, T_{igs} is an indicator variable for random assignment to treatment, and Y_{igs}^{t-1} is math achievement at time $t-1$ (baseline). The parameter of interest is β, which captures the causal effect of the intervention. We fit variations of this model that interacted the treatment dummy with students' grade, sex, and baseline achievement (continuous or by within-grade quartile) to understand whether the intervention was more helpful for some sub-groups of students. We also interacted 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. We pre-specified these analyses in the American Economics Association's Trial Registry (RCT ID: AEARCTR-0002459). Finally, as mentioned in the prior section, we also use IPW estimates and Lee (2009) bounds to show that attrition does not alter our general pattern of results.

5. Results

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), but it reflects the constraints that schools face to integrate this software into their regular instruction (e.g., availability of classrooms and computers, coordination between teachers' timetables, time lost by taking students to the computer lab) (see, e.g., Ferman et al., 2019; Rodriguez-Segura, 2021).

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 computer-based individualized instruction on math achievement, which could be improved upon if schools increased and sustained the use of the software.

For students who were exposed to the software, the diagnostic assessment confirmed that they had a clear need for individualized instruction. 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 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 individualized 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 software's individualized instruction 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 diagnosed to be at a grade 3 level saw grade 3 exercises, see Figure 3).[4] 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; 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 on numbers (95% of total exercises attempted), and much less on geometry (4.5%) or data (0.5%, Table A.1).[4] 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.

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 2 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 computer-based individualized instruction 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 computer-based individualized instruction.[5]

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 ITT effect remained virtually unchanged if we weighted results by the inverse

probability of each student's participation 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).

Heterogeneous effects on math achievement

The null average effects, however, masked important heterogeneous impacts. We investigated whether the effect of computer-based individualized instruction 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 this indicator with indicator variables for each student's within-grade quartile of baseline achievement (Table 4, columns 1 and 2, where the first row reflects effects on the bottom percentile and bottom quartile, respectively).[6] 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).

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, column 3, where the first row reflects effects on the bottom percentile), and with the

difference between each student's enrolled and diagnosed grade level (Table 4, column 4, where the first row reflects effects on students who were more than three grade levels behind). Both specifications indicated that students with lower math achievement at baseline saw larger treatment effects than their peers (as suggested by the negative 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), most 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.

Average effects on interaction with CAL software

Given that the individualized instruction feature of the software may assign students to lower-grade and/or remedial exercises, it is possible that it leads them to potentially completing fewer units in the CAL platform than control students. This would be problematic because, while we expect that computer-based individualized instruction would *positively* impact the math achievement of treatment students, we would also expect that completing fewer units would *negatively* impact their achievement, and the average effect that we estimate may confound these conflicting influences.

We addressed this possibility in three ways in Table A.5. First, we estimated the effect of the intervention on the number of sessions completed on the CAL platform. Treatment students spent less than 1% more sessions than control students, but the difference between the two is statistically insignificant. Second, we estimated the effect of the intervention on the total time spent on the platform. Treatment students spent 2.8% more minutes than control students, but again, the difference was not statistically significant. Third, we estimated the effect of the intervention on the total time spent on the platform, holding the number of sessions completed constant (to estimate the effect of the intervention on time spent per session). Per session, treatment students spent 2.3% more minutes on the platform, but the difference was statistically insignificant. In short, we did not see any compelling evidence that the individualized instruction feature of the software held treatment students back.

6. Conclusion

This paper presents one of the first studies to isolate the effect of computer-based individualized instruction 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 their grade's 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 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. These results suggest that technology-enabled individualized instruction 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.

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 individualized instruction feature in the Mindspark software, which was found effective in both its after-school (Muralidharan et al., 2019) and in-school formats (Muralidharan & Singh, 2019), may play an important part in improving learning for low performers. This finding is intuitive, but to our knowledge, there are no experimental studies in developing countries that are designed to isolate the effects of individualized instruction 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 individualized instruction as well as many other features (for reviews, see Bulman & Fairlie, 2016; Tauson & Stannard, 2018; Escueta et al., 2020; Rodriguez-Segura, 2021). The present study is a much needed step in identifying the *features* that makes some products effective—especially, for low performers.

Second, our study also contributes to the growing evidence of differentiated learning more 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., 2007; Banerjee et al., 2010)—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 individualized instruction.

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 students already have higher mean achievement and lower variability in that achievement, and where schools have more resources to provide students with more opportunities to improve, as it may be the case with the model public schools in our study. Yet, equally importantly, these interventions may improve the performance of low achievers in these settings, who may not have reached a performance level that allows them to reap the benefits of better peers and resources.

Finally, our study demonstrates how to leverage the increasing prevalence of educational software products to run rapid-cycle randomized evaluations that shed light on the merits of intuitively appealing yet largely untested educational strategies. First and foremost, we show that it is possible to understand the relative contributions of individual components of effective

"bundled" software interventions (see Muralidharan, 2017) by temporarily deactivating them to evaluate their contribution. Importantly, this approach does not require recruiting additional users and it yields an estimate of the impact of each component in a brief period, lowering the costs for experimentation for the software developers. 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 to improve their effectiveness (Ganimian et al., 2020). More broadly, we also illustrate how these evaluations can yield broader lessons for pedagogy by documenting the need for and the impact of specific pedagogical strategies (in this case, individualized instruction). Notably, evaluating strategies using technology can offer valuable information on important mechanisms (in this case, how the material with which students interact becomes closer to their ability) and on effects for target sub-groups (in this case, initially low performing students), which is more challenging to do in evaluations of in-person pedagogical interventions.

We believe further research on technology-enabled individualized instruction can complement the present study on multiple fronts. First, it is important for the field to understand whether coarser types of computer-based individualized instruction, which are far more prevalent than the unique feature in the Mindspark software, are as effective at improving student learning. While many studies claim to be evaluating software products that include some type of individualized instruction, the bulk of these products either require students to attain a certain level of proficiency before moving on to the next unit or simply decrease the difficulty of the activities, with little regard for the specific gaps in students' knowledge, let alone their underlying misconceptions (for a review of product features, see Appendix C of Muralidharan et al., 2019). We suspect that this distinction, which is often glossed over in discussions of the

effectiveness of educational software products, may make a meaningful difference in improving students' learning outcomes. Given that not all software developers can count on the resources or years of iteration as Educational Initiatives, the developer of Mindspark, it would be useful to understand whether more attainable individualized instruction features can confer similar benefits.

Second, our study illustrates the importance of testing the effectiveness of technology-enabled individualized instruction across school systems and school types within them. Specifically, we expect the potential contribution of computer-based individualized instruction to be larger in countries like India, where there is vast heterogeneity in students' preparation for school (ASER, 2021). This heterogeneity is also present in other South Asian countries and Sub-Saharan African nations (see, e.g., Uwezo, 2015, 2016, 2019; ASER Pakistan, 2020), where individualized instruction likely holds similar promise, but less so in other low- and middle-income countries, and it seems crucial to understand the extent to which the impact of individualized instruction hinges on this degree of variability in achievement. We also expect the margin for impact of computer-based individualized instruction to be larger in traditional public schools, which typically have students with lower and more variable achievement than the public model schools in which we conducted our study. In short, understanding whether technology-enabled individualized instruction is a stopgap measure to deal with very high levels of heterogeneity in student preparation, or a more generally useful pedagogical approach, seems like a first-order priority for ongoing research on education technology in the developing world.

References

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

Angrist, J., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal, 112*(482), 735-765. doi:10.1111/1468-0297.00068

Angrist, N., de Barros, A., Bhula, R., Chakera, S., Cummiskey, C., DeStefano, J., Floretta, J., Kaffenberger, M., Piper, B., & Stern, J. (2021). Building back better to avert a learning catastrophe: Estimating learning loss from COVID-19 school shutdowns in Africa and facilitating short-term and long-term learning recovery. *International Journal of Educational Development, 84*, 102397. Retrieved from https://shared.rti.org/content/calculating-educational-impact-covid-19-part-ii-using-data-successive-grades-estimate

Araya, R., Arias Ortiz, E., Bottan, N. L., & Cristia, J. P. (2019). *Does gamification in education work?: Experimental evidence from Chile*. (IDB Working Paper No. IDB-WP-982). Inter-American Development Bank. Washington, DC.

ASER. (2018). *Annual status of education report (ASER) 2018: Provisional*. ASER Centre. New Delhi, India.

ASER. (2021). *Annual status of education report (ASER) 2021: Provisional*. ASER Centre. New Delhi, India.

ASER Pakistan. (2020). *Annual Status of Education Report (ASER) Pakistan 2019: Provisional*. ASER Pakistan Secretariat. Lahore, Pakistan.

Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K., & Iqbal, S. A. (2020). Simulating the potential impacts of Covid-19 school closures on schooling and learning outcomes: A set of global estimates. *The World Bank Research Observer, 36*(1), 1-40.

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof to concept to scalable policies: Challenges and solutions, with an application *Journal of Economic Perspectives, 31*(4), 73-102.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy, 2*, 1-30. doi:10.1257/pol.2.1.1

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics, 122*, 1235-1264. doi:10.1162/qjec.122.3.1235

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia*. (Impact Evaluation Series No. 29). The World Bank. Washington, DC.

Berlinski, S., & Busso, M. (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economics Letters, 156*, 172-175.

Beuermann, D. W., Cristia, J., Cruz-Aguayo, Y., Cueto, S., & Malamud, O. (2015). Home computers and child outcomes: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics, 7*(2), 53-80.

Beuermann, D. W., Naslund-Hadley, E., Ruprah, I. J., & Thompson, J. (2013). The pedagogy of science and environment: Experimental evidence from Peru. *The Journal of Development Studies, 49*(5), 719-736. doi:10.1080/00220388.2012.754432

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

Bulman, G., & Fairlie, R. W. (2016). *Technology and education: Computers, software, and the Internet*. (NBER Working Paper No. 22237). National Bureau of Economic Research (NBER). Cambridge, MA.

Carrillo, P., Onofa, M., & Ponce, J. (2011). *Information technology and student achievement: Evidence from a randomized experiment in Ecuador*. (IDB Working Paper No. IDB-WP-223). Inter-American Development Bank. Washington, DC.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2017). Technology and child development: Evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics, 9*(3), 295-320.

Das, J., & Zajonc, T. (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., Ganimian, A. J., & Venkatachalam, A. (2022). Which students benefit from independent practice? Experimental evidence from a math software in private schools in India. *Journal of Research on Educational Effectiveness, 15*(2), 279-301.

Djaker, S., Ganimian, A. J., & Sabarwal, S. (2022). *Out of sight, out of mind? The gap between students' performance and teachers' estimations in Bangladesh and its implications for instruction*. Unpublished manuscript. Steinhardt School of Culture, Education, and Human Development, New York University. New York, NY.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739-1774. doi:10.1257/aer.101.5.1739

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature, 58*(4), 897-996.

Fairlie, R. W., & Robinson, J. (2013). Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *American Economic Journal: Applied Economics, 5*(3), 211-240.

Ferman, B., Finamor, L., & Lima, L. (2019). *Are public schools ready to integrate ed-tech classes with Khan Academy?* (MPRA Paper No. 94736). Munich Personal RePEc Archive (MPRA). Munich, Bavaria.

Ganimian, A. J., & Djaker, S. (2022). *How can developing countries address heterogeneity in students' preparation for school? A review of the challenge and potential solutions*. Unpublished manuscript. Steinhardt School of Culture, Education, and Human Development, New York University. New York, NY.

Ganimian, A. J., Hess, F. M., & Vegas, E. (2020). *Realizing the promise: How can education technology improve learning for all?* Brookings Institution. Washington, DC.

Ganimian, A. J., Muralidharan, K., & Walters, C. R. (2022). *Augmenting state capacity for child development: Experimental evidence from India*. (NBER Working Paper No. 28780). National Bureau of Economic Research (NBER). Cambridge, MA.

IEA. (2017). *TIMSS 2019: Assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Kaffenberger, M. (2020). *Modeling the long-run learning impact of the COVID-19 learning shock: Actions to (more than) mitigate loss*. RISE Insights. Research on Improving Systems of Education (RISE).

Kaffenberger, M., & Pritchett, L. (2020). *Failing to plan? Estimating the impact of achieving schooling goals on cohort learning*. (RISE Working Paper No. RISE-WP-20/038). Research on Improving Systems of Education (RISE). London, UK.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Kumar, N. (2020). *Public school quality and student outcomes: Evidence from model schools in India*. Unpublished manuscript. University of California, San Diego. San Diego, CA.

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review, 47*, 34-48. doi:10.1016/j.econedurev.2015.03.005

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M., & Rozelle, S. (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. doi:10.1080/19439342.2013.780089

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (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. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies, 76*, 1071-1101.

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

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

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

Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics, 126*, 987-1027. doi:10.1093/qje/qjr008

MHA. (2012). *15th Census of India*. 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. New Delhi, Delhi.

MIA. (2015). *Medición independiente de aprendizajes: Porque la educación es de todos, la responsabilidad es mía*. Medición Independiente de Aprendizajes (MIA). Ciudad de México, México.

MLJ. (2009). *The right of children to free and compulsory education act, 2009*. Ministry of Law and Justice, Government of India. New Delhi, Delhi.

Mo, D., Bai, Y., Boswell, M., & Rozelle, S. (2016). *Evaluating the effectiveness of computers as tutors in China*. (3ie Impact Evaluation Report No. 41). International Initiative for Impact Evaluation (3ie). New Delhi, India.

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

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., & Rozelle, S. (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. doi:10.1080/19439342.2014.911770

Mo, D., Zhang, L., Wang, J., Huang, W., Shi, Y., Boswell, M., & Rozelle, S. (2015). Persistence of learning gains from computer assisted learning: Experimental evidence from China. *Journal of Computer Assisted Learning, 31*, 562-581.

Muralidharan, K. (2017). Field experiments in education in developing countries. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments (Vol. 1)*: North Holland.

Muralidharan, K., & Singh, A. (2019). *Improving schooling productivity through computer-aided personalization: Experimental evidence from Rajasthan*. Unpublished manuscript. University of California, San Diego. San Diego, CA.

Muralidharan, K., & Singh, A. (2020). *Improving school productivity through computer-aided instruction: Experimental evidence from Rajasthan*. Unpublished manuscript. University of California, San Diego. San Diego, CA.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review, 109*(4), 1426-1460.

NIEPA. (2018). *U-DISE flash statistics 2016-17*. National Institute of Educational Planning and Administration (NIEPA). New Delhi, India.

Pritchett, L., & Beatty, A. (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. (2021). Educational technology in developing countries: A review of the evidence. *The World Bank Research Observer*.

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

Sinha, S., Banerji, R., & Wadhwa, W. (2016). *Teacher performance in Bihar, India: Implications for education*. Washington, DC: The World Bank.

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

Uwezo. (2015). *2014 Tanzania annual assessment report*. Uwezo East Africa Regional Office. Nairobi, Kenya.

Uwezo. (2016). *2015 Kenya annual assessment report*. Uwezo East Africa Regional Office. Nairobi, Kenya.

Uwezo. (2019). *Are our children learning? Uwezo Uganda eighth learning assessment report*. Twaweza East Africa. Kampala, Uganda.

Van Klaveren, C., Vonk, S., & Cornelisz, I. (2017). The effect of adaptive versus static practicing on student learning-evidence from a randomized field experiment. *Economics of Education Review, 58*, 175-187.

World Bank. (2016a). *Time on task study in secondary schools: Madhya Pradesh & Tamil Nadu, 2015-2016*. Unpublished manuscript. The World Bank and Educational Initiatives. New Delhi, India.

World Bank. (2016b). *What is happening inside classrooms in Indian secondary schools? A time on task study in Madhya Pradesh and Tamil Nadu*. The World Bank. Washington, DC.
Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)*. Westport, CT: American Council on Education and Praeger Publishers.

Endnotes

[1] We initially planned to introduce two more treatment groups. Yet, due to technical difficulties, we abandoned them shortly after baseline and excluded students in those experimental groups from the present study. These interventions are described in our pre-registration plan: https://www.socialscienceregistry.org/trials/2459.

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

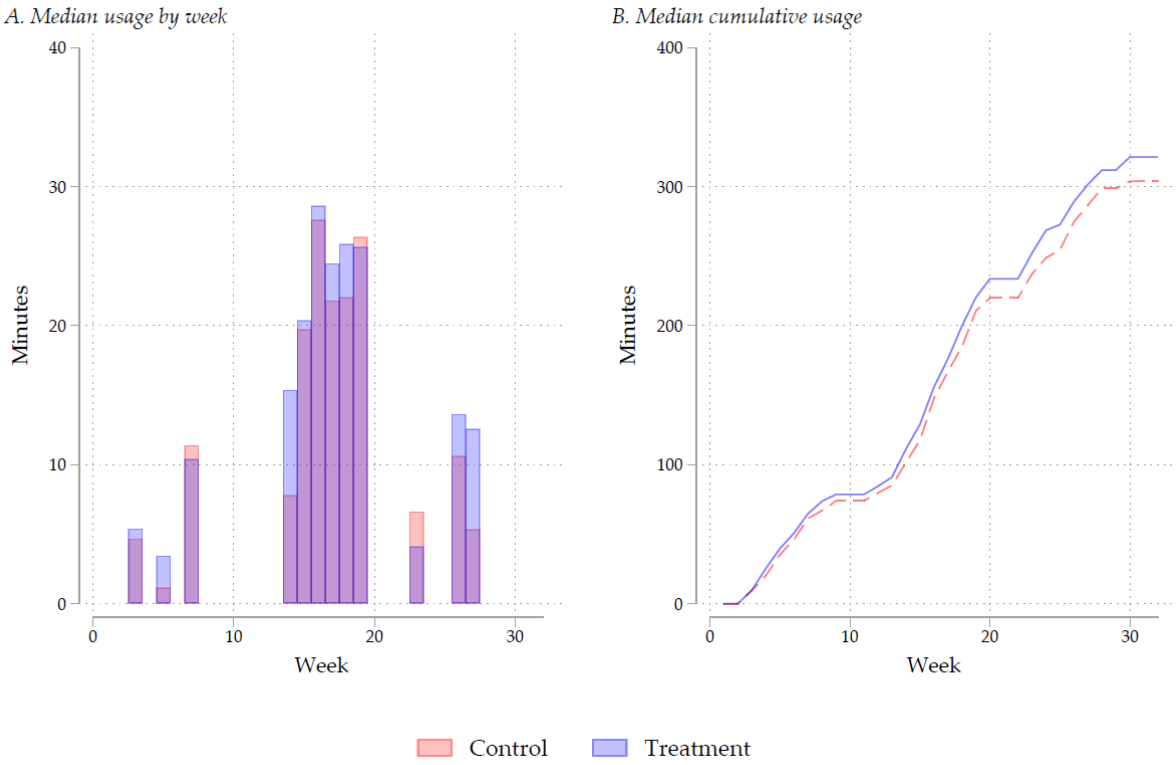
[3] We have no way of knowing why the material for treatment students enrolled in some grades (e.g., grade 6) were concentrated on a specific lower grade (e.g., grade 4), while the material for other grades (e.g., grade 8) were spread across several lower grades (e.g., grades 4 to 8). It is possible that most students enrolled in lower grades needed reinforcement of basic content, whereas those enrolled in higher grades had more diverse needs.

[4] 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.

[5] We do find a statistically significant effect of computer-based individualized instruction 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.

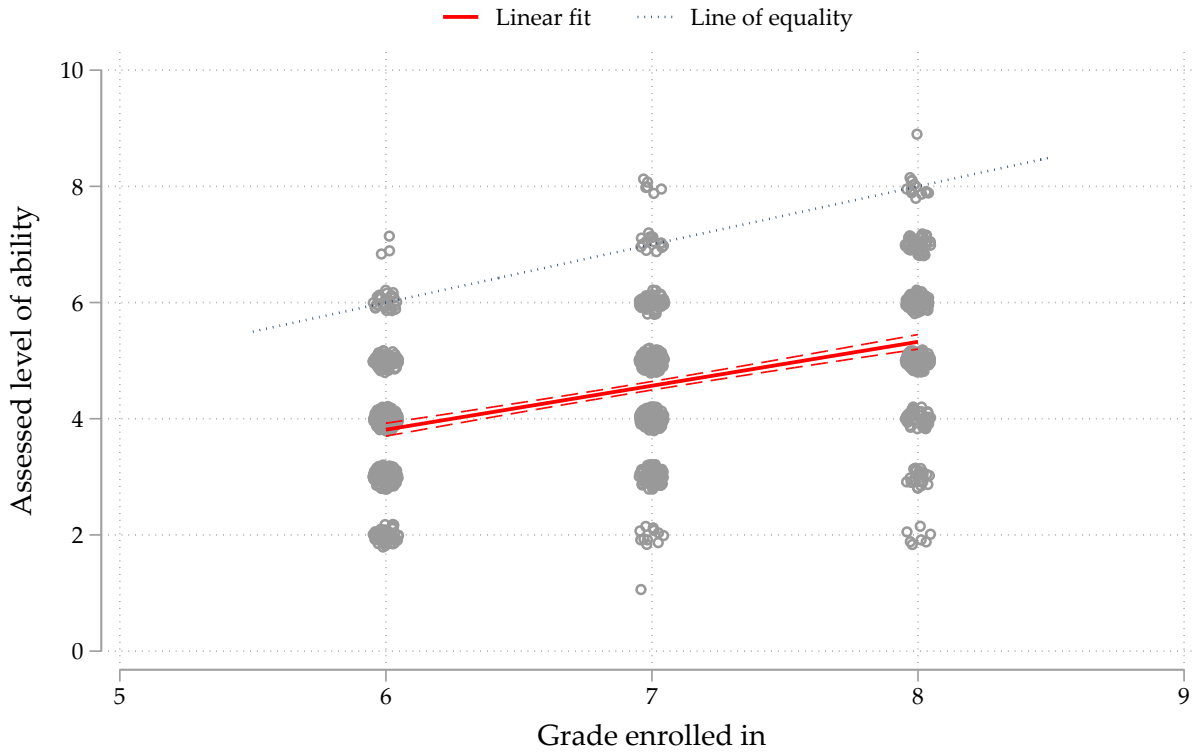
[6] We report on impacts by within-grade quartile. The intent-to-treat effect for the within-grade bottom *half* of students is 0.18 SDs ($p<0.01$). The respective effect for the top half of students is -0.05 SDs ($p>0.1$).

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



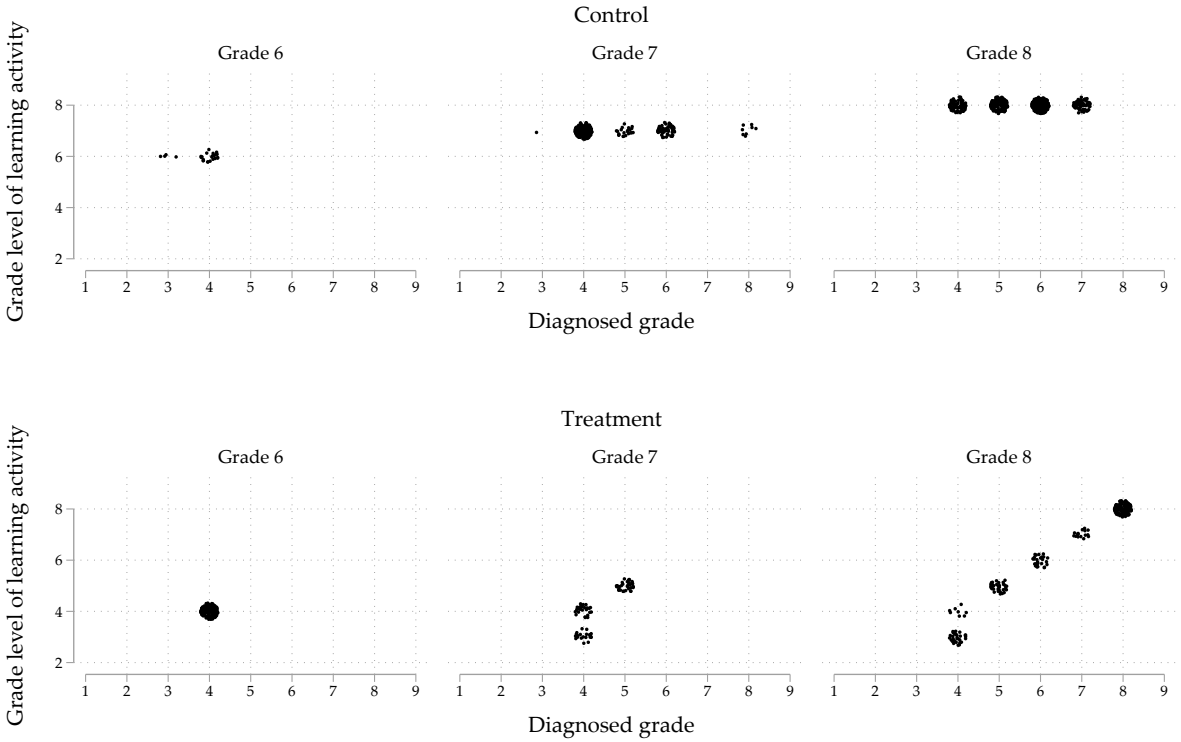
Notes: This figure shows the weekly (panel A) and cumulative (panel B) usage of the CAL platform for the median student, by experimental group. This figure includes all students observed at baseline and endline, regardless of whether they used the software (99.2% of students did). Usage is binned by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

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



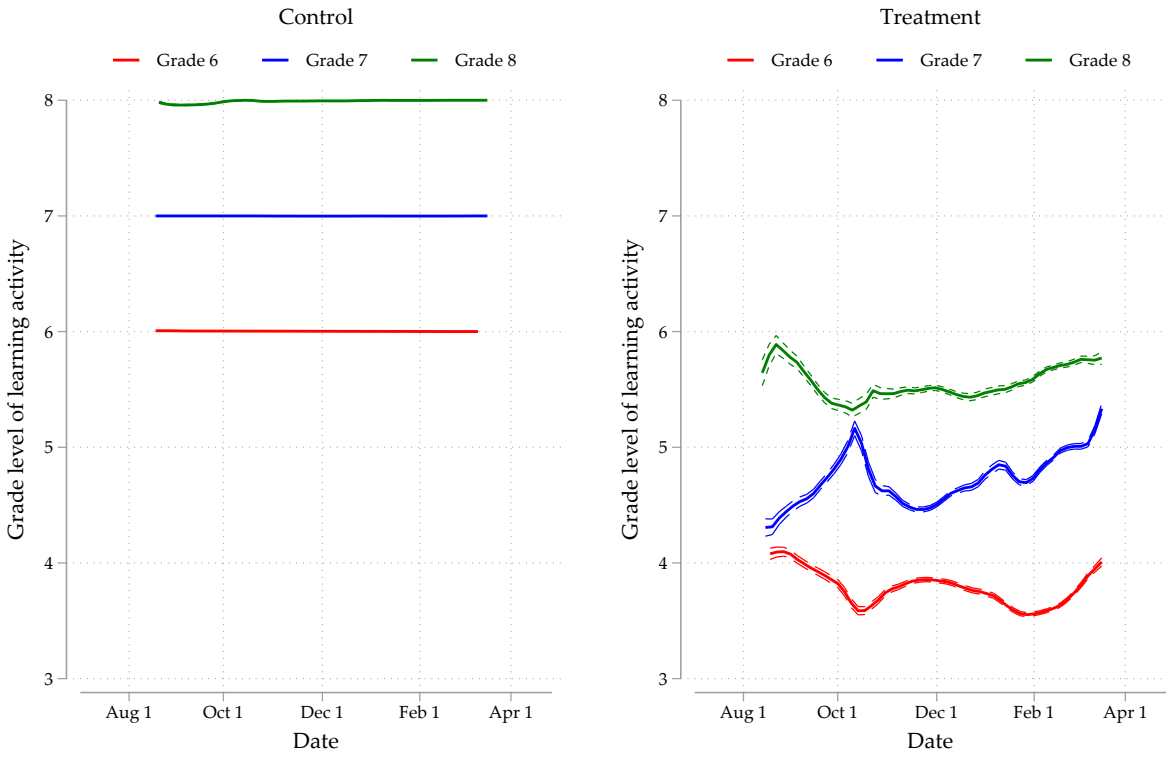
Notes: This figure shows the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. These data are from the initial diagnostic test and do not reflect any instruction provided by Mindspark. We find a general deficit between average 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



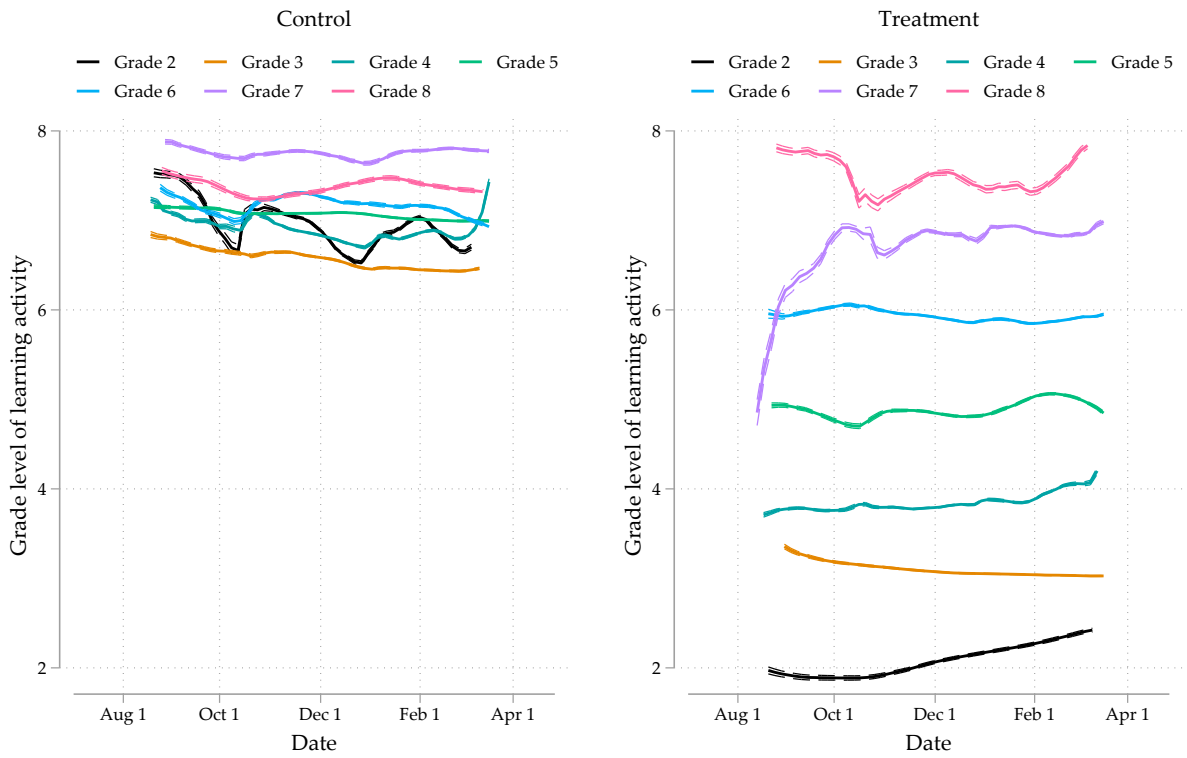
Notes: This figure shows, by treatment group, the grade level of learning activities administered by the computer adaptive system to students, on a single day (shortly after activating the study, on August 30, 2017). For simplicity, the figure omits exercises that are also mapped to another, adjacent 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 computer-based individualization of content, by enrolled grade and experimental group



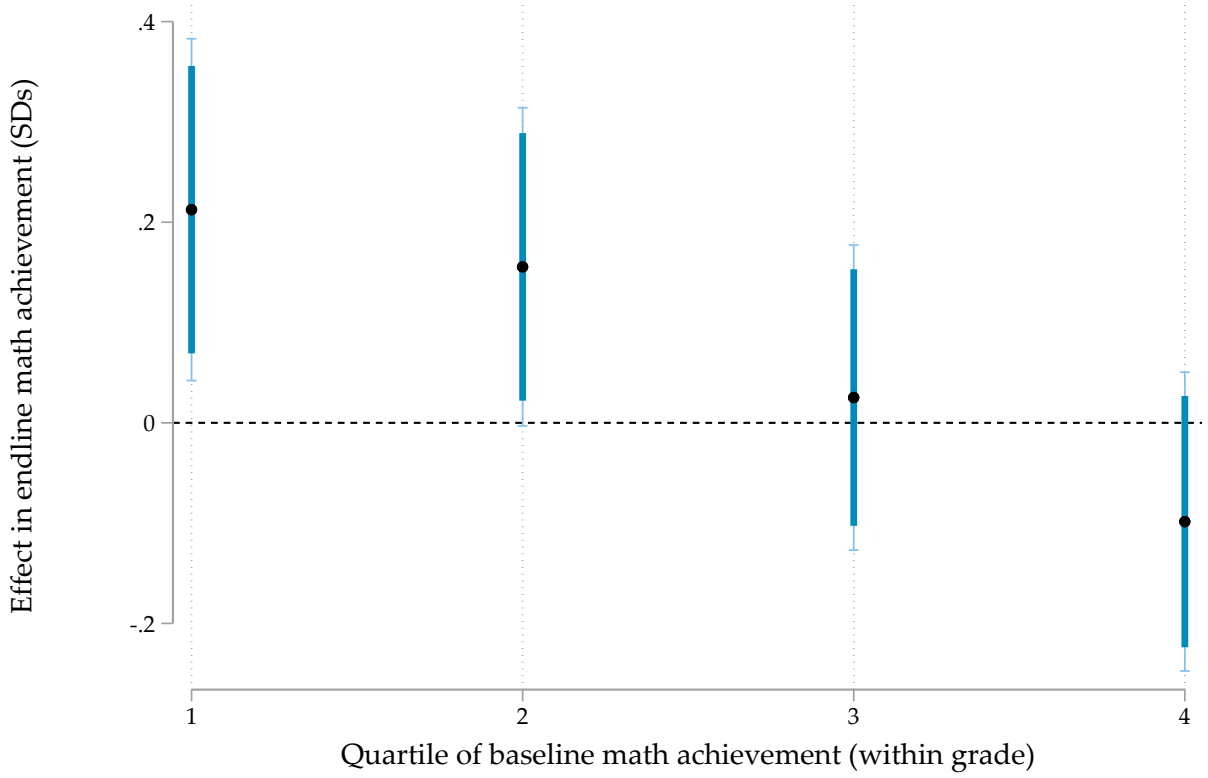
Notes: This figure shows, by experimental group, kernel-weighted local mean smoothed lines relating the level of difficulty of the exercises attempted by students with the date of administration. Separate lines reflect the actual grade of enrolment. The software was activated on August 6, 2017, but its first usage was registered on August 10, 2017. For simplicity, the figure omits learning activities that 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 computer-based individualization of content, by diagnosed grade and experimental group



Notes: This figure shows, by experimental 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 that 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 computer-based individualized instruction 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 computer-based individualized instruction 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
Baseline score (C)	0.080	0.039	0.080	0.087	0.087
	[0.986]	[0.995]	[0.986]	[0.984]	[0.984]
Baseline score (T)	0.051	-0.024	0.051	0.053	0.053
	[0.997]	[1.110]	[0.997]	[0.998]	[0.998]
Endline score (C)	0.228	0.191	0.228	0.232	0.232
	[0.993]	[1.021]	[0.993]	[0.995]	[0.995]
Endline score (T)	0.255	0.210	0.255	0.259	0.259
	[0.967]	[1.081]	[0.967]	[0.966]	[0.966]
Growth (C)	0.148***	0.153***	0.148***	0.145***	0.145***
	(0.029)	(0.030)	(0.029)	(0.029)	(0.029)

Notes: This table shows the intent-to-treat (ITT) effect of computer-based individualized instruction 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 student's 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. The table also shows the mean math (IRT-scaled) score for the control group (C) and treatment group (T), respectively, at baseline and endline. The last row shows the mean growth in the control group (the difference between the endline and baseline scores). Standard errors are shown in parentheses; standard deviations are shown in brackets. All estimations of treatment effects include randomization-strata fixed effects. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: ITT effect of computer-based individualized instruction on math achievement at endline, by content and cognitive domain

	(1) Numbers	(2) Geometry	(3) Measurement	(4) Knowing	(5) Applying	(6) Reasoning
A. All students						
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 (std.)	0.122***	0.133***	0.134***	0.118***	0.124***	0.168***
	(0.005)	(0.006)	(0.007)	(0.005)	(0.005)	(0.009)
N (students)	1,078	1,078	1,078	1,078	1,078	1,078
R-squared	0.503	0.465	0.411	0.471	0.534	0.418
FWER-adj. p-value	0.592	0.43	0.239	0.51	0.182	0.38
B. Low performers						
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 (std.)	0.127***	0.136***	0.119***	0.121***	0.133***	0.130***
	(0.013)	(0.016)	(0.019)	(0.013)	(0.013)	(0.023)
N (students)	1,078	1,078	1,078	1,078	1,078	1,078
R-squared	0.515	0.476	0.422	0.480	0.541	0.424
FWER-adj. p-value	0.298	0.85	0.459	0.744	0.068	0.858

Notes: This table shows the intent-to-treat (ITT) effect of computer-based individualized instruction 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%.

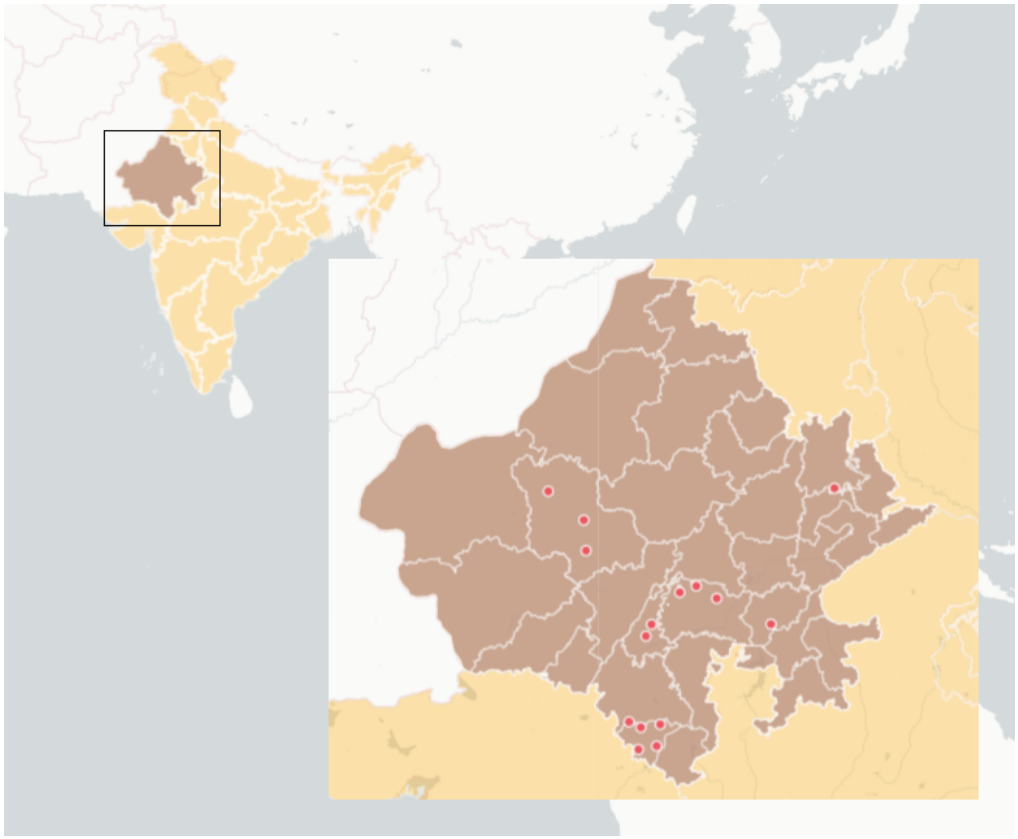
Table 4: Heterogeneous ITT effects on math achievement at endline, by students' baseline performance

	(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 computer-based individualized instruction 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). In column 4, the reference category consists of students who are more than three levels behind. 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.5 (i.e., for 16 tests).

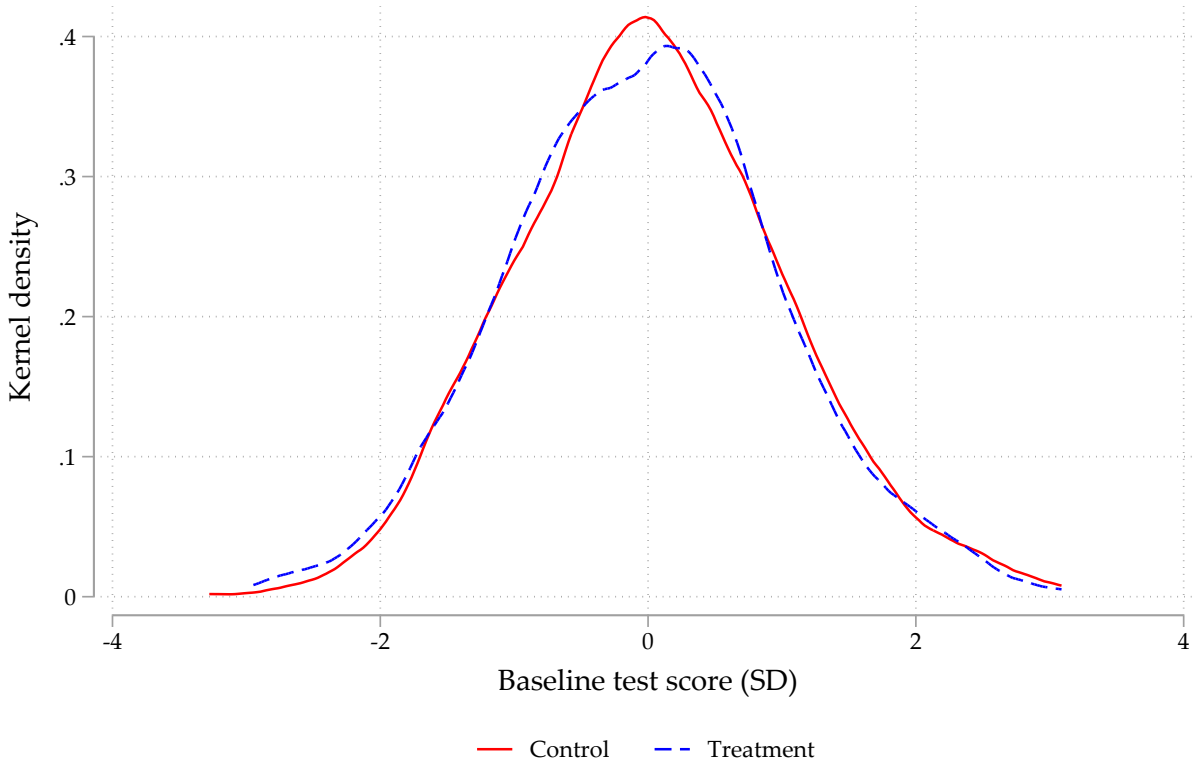
Appendix A: Additional graphs and tables

Figure A.1: Map of study districts and schools



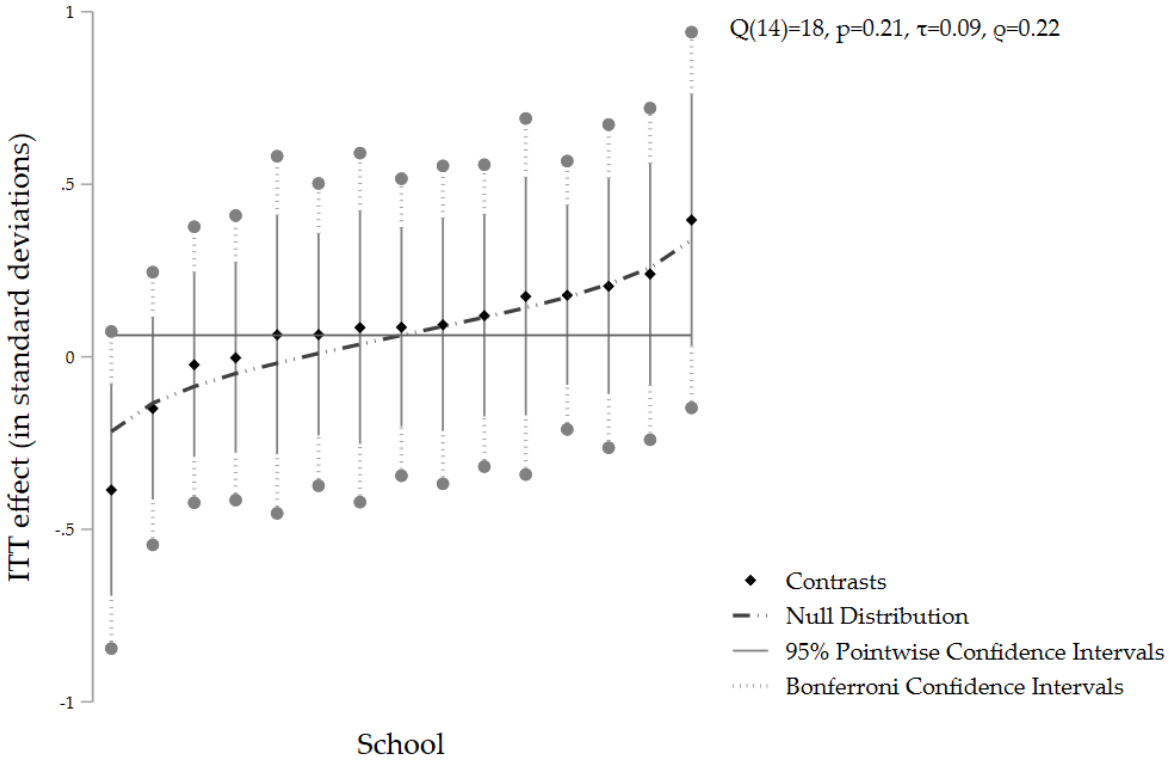
Notes: This figure shows the state of Rajasthan (in brown) and the location of study schools (in red).

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



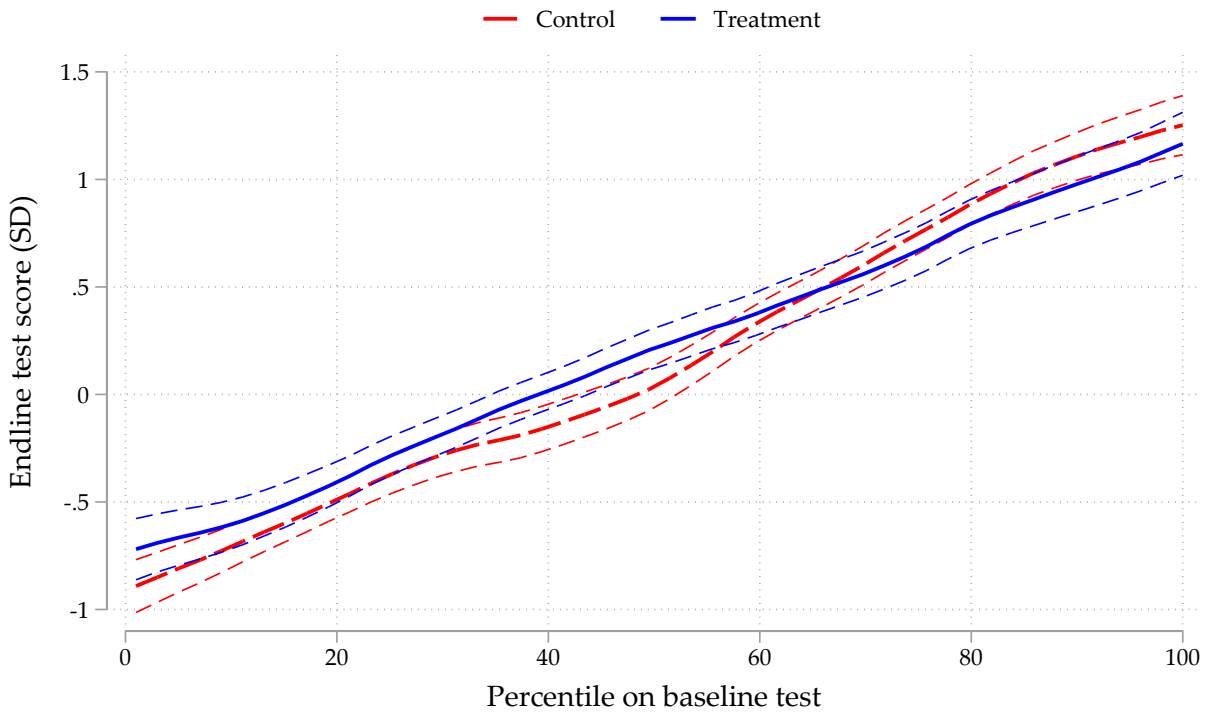
Notes: This figure shows the distribution of scores in the baseline assessment of math for control and treatment students. Scores were scaled using a two-parameter logistic Item Response Theory (IRT) model. This figure includes all students present at baseline and at endline.

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



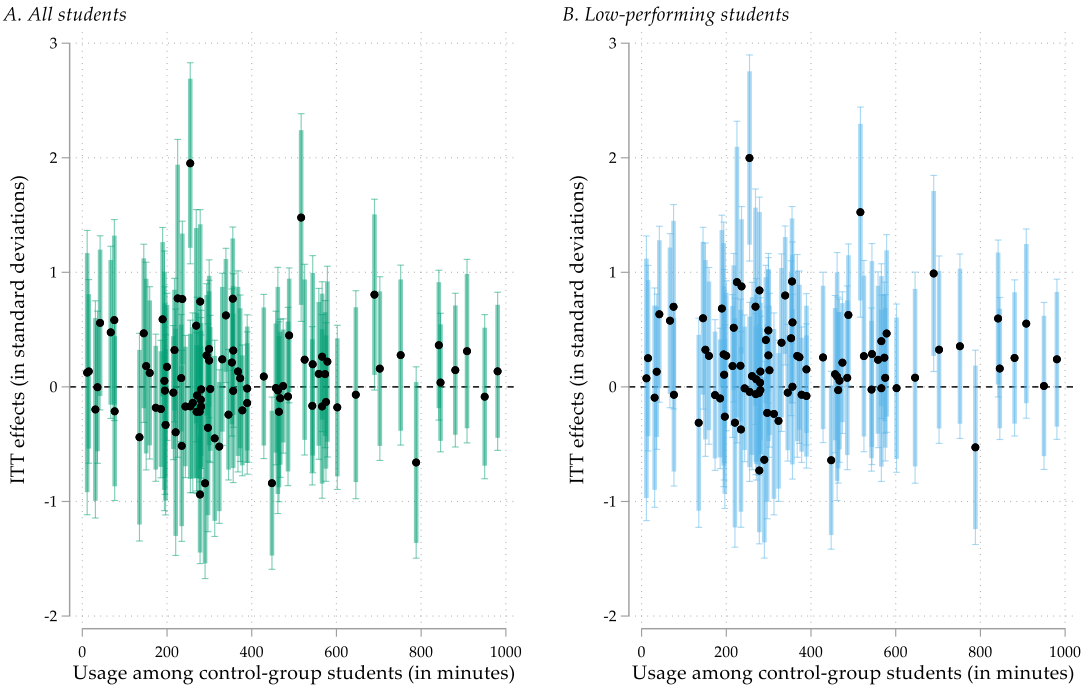
Notes: This figure provides a "caterpillar plot" of ITT effects by school (cf. von Hippel & Bellows, 2018). Each black dot refers to the point estimate for a given school. Bonferroni confidence intervals adjust standard errors for multiple hypothesis testing. The black solid line shows the null distribution 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) baseline percentiles on the baseline test



Notes: The figure presents kernel-weighted local mean smoothed plots that 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 individualized instruction on students' achievement in math at endline (after nine 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 (of 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	4.2%
Ratio and proportion	15	0.04%
Percentages and commercial math	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: The 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 individualized instruction 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

Notes: This table shows the Lee (2009) bounds on the intent-to-treat (ITT) effect of individualized instruction 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 individualized instruction on math achievement at endline, by repeated and non-repeated items

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

Notes: This table shows the intent-to-treat (ITT) effect of individualized instruction on students' achievement in items administered in both baseline and endline (which we call "repeated items" in column 1) and items that were first introduced in the endline (which we call "non-repeated items" in column 2) 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. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Heterogeneous ITT effects of individualized instruction 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 individualized instruction 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 practice exercises on usage of CAL platform

	(1)	(2)	(3)
	Number of sessions completed (log)	Total minutes spent on CAL platform (log)	
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 practice exercises 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 exclude nine students who did not spend any time on the software. * significant at 10%; ** significant at 5%; *** significant at 1%.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies, 76*, 1071-1101.

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

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