Which Students Benefit from Independent Practice? Experimental Evidence from a Math Software in Private Schools in India*

Andreas de Barros†
Alejandro Ganimian‡
Anuja Venkatachalam§

May 23, 2021

Abstract

This study is one of the first evaluations of independent (i.e., self-guided) practice in math in a developing country. We randomly assigned 4,461 students in grades 4-7 in "unaided" private schools across seven Indian cities who were using a computer-assisted learning software to: (a) a control group, in which they moved from one unit to the next upon completion; or (b) a treatment group, in which they had to complete practice exercises before progressing to the next topic. After six months, the additional practice had a precisely estimated null effect on the math achievement of the average student. However, treatment students with low initial performance outperformed their control counterparts by 0.14 standard deviations (SDs). Our results suggest that independent practice may help private-school students from relatively well-off families in need of catching up with their peers.

Keywords: computer-aided learning, India, math instruction, practice exercises, private schools

JEL classification: C93, I21, I22, I25

Study pre-registration: Pre-registered at the AEA RCT Trial Registry (AEARCTR-0002455)

*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. 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

§Data Journalist, DataLEADS. E-mail: anujav04@gmail.com

There is broad consensus among math educators about the importance of practice for primary- and middle-school students (see Woodward et al. 2012). Yet, the type of practice that has been found effective is often preceded by a considerable degree of scaffolding from teachers, either through "task lists" of specific steps (see, e.g., Hohn and Frey 2002; Verschaffel et al. 1999), "self-questioning" checklists (e.g., A. K. Jitendra et al. 1998; Asha K. Jitendra et al. 2011) or processes (Cardelle-Elawar 1990, 1995; King 1991; Kramarski and Mevarech 2003; Mevarech and Kramarski 2003). Far less is known about "independent" (i.e., self-guided) practice that is not necessarily matched to specific metacognitive strategies that provide students with a structure to think about how to approach a type of problem. This type of practice is becoming increasingly important with the advent of coronavirus, which has suddenly required that students take a more independent role in their education.

More specifically, we still know far too little about whether (and if so, how) independent practice affects students at varying levels of preparation for schooling differently. Most of the evaluations of practice-based interventions cited above do not estimate heterogeneous effects by baseline achievement, and many of those that do lack sufficient statistical power to detect them, so it has been challenging to make sense of null effects for interactions between interventions and students' baseline performance. Ex ante, it is not clear what to expect. On the one hand, low-performing children may benefit from learning at their own pace and independent practice may allow them to do so. Accordingly, multiple impact evaluations of software-enabled practice in these contexts have consistently found small-to-moderate positive effects on student achievement (see, e.g., Bettinger et al. 2020; Lai et al. 2013; Lai et al. 2015; Lai et al. 2016; Mo et al. 2015; Pitchford 2015). On the other hand, these children are often several grade levels behind curricular expectations, so additional practice may not be as helpful as remedial work that focuses on building foundational skills and remedying misconceptions. That may be the reason why software that dynamically adjusts to the performance of each student has yielded even larger effects on achievement than those that review the material covered in school during a given week (e.g., A. V. Banerjee et al. 2007a; Muralidharan, Singh, and Ganimian 2019).

This paper presents one of the first studies of the effects of technology-enabled independent practice in a developing country. During the 2017-2018 school year, we partnered with an educational-assessment firm in India to randomly assign 4,461 students in grades 4-7 in private-"unaided" (i.e., independently funded) schools across seven Indian cities that were using a computer-assisted learning (CAL) software to: (a) a control group, in which students complete a set of "learning exercises" (which present new concepts) on a given topic in math (e.g., operations with fractions) and then move on to the next topic; or (b) a treatment group, in which students also complete the learning exercises but are then asked to complete a set of "practice exercises" (which seek to build procedural knowledge) before moving on to the next topic. Specifically, given that practice exercises were a feature of the software, we deactivated them for control-group students

during the study period, building on prior studies that infer the effect of a policy or program by examining the consequences of its discontinuation (Argys et al. 2020; Chyn 2018; Dynarski 2003; Fishman et al. 2017; Nakajima 2020). We discuss the implications of this approach for our research question in greater detail below.

We can verify that the intervention was implemented mostly as intended: during the six months of the study, the median student across both experimental groups interacted with the CAL software for more than 700 minutes and treatment students interacted with the practice exercises for more than 60 minutes. Students could interact with the software during or after school, but the bulk of usage occurred at home (66% on average during the study), which makes it particularly relevant for the current context of the pandemic. The three topics in which students attempted most practice exercises were: measurement (14% of all exercises attempted), fractions (10%), and numbers (9.1%). Further, interaction with the software remained relatively constant during the study—typically, students used the software between 20 and 40 minutes per week. Treatment students collectively attempted nearly half a million practice exercises.

We report two main sets of results. First, after six months, we found that practice exercises had a precisely estimated null effect on math achievement, as measured by an independent test designed by the research team (not by the software developers): the average treatment student performed 0.01 standard deviations (SDs) better than their counterpart in the control group, and the difference was statistically insignificant. We can rule out positive effects larger than 0.06 SDs (as per the upper limit of the estimate's 95% confidence interval). In fact, we observe effects consistently estimated around zero across all topics and skills in the assessment. Further, we do not see any relationship between the number of practice exercises completed by the average student and his/her achievement in math.

Yet, these average impacts mask non-trivial heterogeneous effects. Initially low-performing students (i.e., those in the bottom quartile of the within-grade baseline math achievement distribution) outperformed their peers in the control group by 0.14 SDs ($p < 0.01$). In fact, they improved in two of the three content domains (numbers and geometry) and all three cognitive domains (knowing, applying, and reasoning) assessed in the endline test. We examine whether low performers who spend more time on the software and do more practice exercises fare better, but to our surprise, we do not find that this is the case. We do not find any statistically significant heterogeneous effects by students' school, grade, or sex.

This study makes several key contributions to research on student learning in developing countries. First, it draws attention to the issue of heterogeneity in the benefits of independent practice. Prior studies in math pedagogy have sought to recruit students from disadvantaged backgrounds to understand the impact of independent practice on a segment of the student population that stands to benefit from additional exposure to the material. Yet, for the most part, this literature relies on

small samples, which prevent researchers from comparing the effect of independent practice across different types of students. The sudden disruption imposed by coronavirus on school systems make this question particularly timely.

Second, this study also contributes to the growing literature on private schools in India. Prior studies have established that these schools, which serve about one in two children in the country (Central Square Foundation 2020), outperform public schools (Muralidharan and Kremer 2008), their advantage is mostly due to student selection rather than better instruction ((Singh 2015), but their lower teacher salaries make them more efficient (Muralidharan and Sundararaman 2015). Our study adds to this growing literature by identifying an intervention that may advance the achievement of this large (and growing) segment of the Indian school system.

More generally, our study illustrates the potential of leveraging CAL software for rapid-cycle randomized experiments that can shed light on the pedagogical approaches that work best. Often, decisions about software features are made by education specialists working alongside developers, informed by research based on small, convenience samples in the United States. In a few cases, it is also guided by so-called "A/B testing", in which a specific feature of the software is offered to a random subset of its users. Yet, as our paper demonstrates, partnerships between researchers and software developers can produce studies that address questions of broader interest to the field of education, ensure that minimal conditions for causal inference are met (e.g., equivalence at baseline and post-attrition), and that the results are shared (and scrutinized) by the relevant scientific community.

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

1 Experiment

1.1 Context

Our study focuses on private schools in India, which have been considerably under-studied given their relative share of domestic and global enrollment in elementary education. They serve 120 million (i.e., nearly one of every two) of all Indian students (Central Square Foundation 2020). If they were considered independently, they would make up the third-largest school system in the world, after the public-school systems of China and India (UIS-UNESCO 2019). Private-school enrollment has increased 33 times between 1973 and 2017, and private-unaided schools, where we conducted our study, accounted for the lion's share of this increase, having surged from 3.4% to 34.8% of total enrollment (Central Square Foundation 2020). According to the latest available data,

these schools enroll 36% and 31% of primary school (i.e., grades 1 to 5) and upper-primary school (grades 6 to 8) students, respectively (NIEPA 2019). While most of them charge relatively low fees (the monthly median fee in an elementary private-unaided school is INR 958 or USD 13 in urban areas and INR 500 or USD 7 in rural areas), they educate one of every three and one of every two students in the top two income quintiles, respectively, and thus enjoy a privileged role in shaping the education of the country's elites (Central Square Foundation 2020).

The nine private-unaided schools in our study spanned seven Indian cities, including: Ahmedabad and Rajkot (in the state of Gujarat), Faridabad (Haryana), Ghaziabad (Uttar Pradesh), Kolkata (West Bengal), New Delhi (Delhi), and Tiruchirappalli (Tamil Nadu). In the 2016-2017 school year, these cities had 13,025 schools and over 2.9 million students (NIEPA 2018a, 2018b), or about the same number of students as the six largest school districts in the U.S. (NCES 2019). Table A1 in Appendix A shows, in all study states except West Bengal, private-unaided schools account for more than a third of enrollment of students at the elementary-school level, and in all states, this share has increased between 35 and 156% in the past decade alone.

Learning-outcomes data for private schools in India is not systematically collected. Yet, existing data suggest that these schools have similar or higher achievement than government (i.e., public) schools. According to the latest (pre-COVID-19) round of a representative assessment of rural schools, only 44% of grade 5 students in government schools could read a grade 2 text, compared to 65% of their private-school counterparts, and only 23% of fifth graders in the public sector could solve a division of a three-digit number by a one-digit number, compared to 40% of their peers in the private sector (ASER 2019). Similarly, in the latest installment of the National Achievement Survey (NAS), which includes private schools affiliated to state boards, students of private-unaided schools slightly outperformed those in private-aided schools in all subjects (math, natural and social sciences, modern Indian languages, and English) (NCERT 2017). These differences are at least partly explained by self-selection into each sector (see, e.g., Muralidharan and Kremer 2008; Muralidharan and Sundararaman 2011; Singh 2015).

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 practice exercises (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 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 Foundation. We were co-principal investigators and research associates on this project.

1.2 Sample

The sample for the study included 4,461 students from grades 4 to 7 across nine private schools in the aforementioned cities. We drew a convenience sample of schools because this was the first study of the Learning Lab and we needed a group of schools that met all of the hardware and software requirements to deliver the CAL software during the study period without major disruptions. We invited 12 private schools to participate. Nine of those schools agreed to participate. We sought informed consent from principals and teachers at those schools.

Students were considered to have attrited from the study if they were absent from school on the day in which the endline assessment was administered or if they dropped out of school. Attrition from the study was low: 4,001 of the 4,461 students who took the baseline assessment (90%) also took the endline assessment. We do not know the share of attritors who dropped out of school, but we can estimate this number by taking advantage of the fact that access to the CAL software was provided through their school. Specifically, for any given week, we can calculate the percentage of students who stopped using the software among attritors and non-attritors. This approach suggests that only a small share of students dropped out of school, but that they were more likely to be attritors: in week 20 (i.e., roughly a month before the endline), 5.2% of attritors had stopped using the software, compared to 0.6% of non-attritors. We found a marginally statistically significant difference in attrition by experimental group: 11% of control students and 9.4% of treatment students attrited, and the difference was significant at the 10% level ($p = 0.052$). We discuss whether this small difference affected our results by reporting inverse-probability weighted (IPW) estimates and Lee (2009) bounds in the Results section.

1.3 Randomization

We randomly assigned the 4,461 students in our sample to: (a) a control group, in which they first completed a set of "learning exercises" (which presented new concepts) on a given topic in math (e.g., operations with fractions) and then moved on to the next topic; or (b) a treatment group, in which students also completed the learning exercises but were then required to complete a set of "practice exercises" (which seek to build procedural knowledge and fluency) before moving on to the next topic. 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 our randomization within each school-by-grade-by-section combination and by students' within-class performance on the CAL platform prior to the start of the study (e.g., one lottery included students in school 1, in grade 4, and section "A", who ranked in this section's top half as per students' prior interaction with the platform).

The randomization of students within the same classroom maximizes statistical power, but its main drawback is that it allows for "spillovers." 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, control-treatment comparisons could under-estimate the effect of the intervention on the math achievement of the former (if they transferred some of their knowledge to the latter). In practice, we believe such spillovers are unlikely for three reasons. First, for practice to meaningfully impact learning, it typically requires more support than a fourth-to-seventh grader could provide to his/her classmates (see, e.g., Palinscar and Brown 1984; Scardamalia, Bereiter, and Steinbach 1984). This is why, as we state in the Introduction, prior studies of practice have focused on initiatives in which teachers provide modeling, coaching, and scaffolding (see also National Research Council 2000) (see also National Research Council, 2000). Second, as we mention in the Introduction and discuss in greater detail in the Results section, the average student spent two-thirds of his/her time interacting with the CAL software at home (rather than at school), where he/she presumably had fewer opportunities to work with classmates. Third, as we also report in the Results section, we find that the intervention had a 0.14 SD effect on initially low-achieving students—precisely, the subset that would benefit most from group work (and thus, the one most prone to spillovers). Even if there had been spillovers, they would have to be similar on magnitude as the benefits from practice accrued to the individual for them to result in precisely estimated null effects. To our knowledge, there are no studies suggesting that spillovers could offset individual gains.

Control and treatment students were comparable on their baseline achievement and sex, regardless of whether we compare all students present at baseline or only those who also took the endline assessment (i.e., non-attritors, see Table A4). In fact, not just the means, but the distribution of baseline achievement was quite similar across experimental groups (see Figure A1 in Appendix A).

1.4 Intervention

The CAL software in which our experiment was embedded, called "Mindspark," 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, Singh, and Ganimian 2019). In our study, students had access to the in-school version, which is currently used by more than 100,000 students in 300 private schools in India and abroad (including

some Arab states in the Persian Gulf). All students continued to receive regular (i.e., teacher-led) instruction during the study period.

There are two important differences between the private-school version of the software (which we use in this study) and the after-school version (which was previously evaluated) that are worth highlighting to interpret our impact estimates. First, the private-school version mostly presents learning and practice exercises at or above grade level: across treatment students, 58% of all practice exercises they completed during the study were at grade level, and 20% were *above* their enrolled grade level (see Table A2). The after-school version caters to children from disadvantaged schools, and consequently emphasizes material below grade level. Second, the topics and skills covered by the private-school version is determined by each student's teacher, based on the curriculum and any additional practice he/she deems appropriate. The content and cognitive domains of the material presented in the after-school version is largely determined by a diagnostic test, which students take when they first interact with the software. Therefore, our impact estimates do not confound the effect of practice with that of the dynamic adaptation of the material, which is largely shut down in the private-school version of the software.[1] Neither principals nor teachers were provided with diagnostics on student learning until *after* the study, so our estimates do not confound the effects of practice with those of information provision. In fact, principals and teachers were "blind" to treatment condition, so it is extremely unlikely that they engaged in any compensatory behavior to only support the learning of treatment students.

We are not interested in evaluating the impact of the software; instead, we use it to randomly assign students to different levels of independent practice in math. Specifically, control and treatment students interacted with the software in a similar manner (Figure 1). First, students were prompted by the software to select a math topic (e.g., fractions); each topic includes between eight and ten units (e.g., subtraction of fractions). Then, they attempted a set of learning exercises that sought to introduce that unit (on average, students complete 16 fill-in-the-blank and multiple-choice questions per unit). Next, the experience of control and treatment students differed: students in the control group moved on to the next unit after completing the learning exercises, whereas those in the treatment group were required to complete practice exercises to consolidate their understanding of the concepts and procedures taught through the learning exercises.[2]

The learning exercises that all students completed differ from the practice exercises that only treatment students were required to complete in four ways. First, as mentioned above, learning

1. As Table A2 shows, there is *some* degree of dynamic adaptation, but within a much narrower set of grades than in the after-school version evaluated by Muralidharan, Singh, and Ganimian (2019).

2. Schools differed in how they used the CAL software: some of them may have used it as part of graded homework, while others may have used it as an additional learning resource. However, these differences should not bias our impact estimate, given that we stratified our randomization by school-by-grade-by-section combination and performance on the platform prior to the study (see Randomization sub-section). Practice exercises could not have been used for high-stakes purposes, as it was not available for control students and since teachers were "blind" to students' assignment to the treatment condition.

exercises introduced new concepts and procedures, whereas practice exercises focused on helping students develop their procedural knowledge (i.e., knowledge about the algorithms to be followed to solve a specific problem) and fluency (i.e., capacity to solve problems rapidly). Second, learning exercises were untimed, but students were given eight to ten minutes (depending on the topic) to attempt each practice exercise. Students who were unable to complete a practice exercise in the allotted time were allowed to attempt it again (with a reset timer) during their next session (to avoid delaying students' progression from one unit to the next). Third, learning exercises were the same for students at all levels of initial achievement, but practice exercises were categorized in three difficulty levels (low, medium, or high), which were presented sequentially (students who attempted low-difficulty exercises graduated to medium-difficulty exercises, and so on, regardless of whether the exercises were answered correctly).[3] Finally, while all units included learning exercises, not all of them included practice exercises.[4]

Importantly, all schools in our study had prior access to the software, which included the practice exercises. In fact, the median student in our study (across both experimental groups) had attempted 683 practice exercises.[5] We deactivated this feature among control students during the evaluation to evaluate its effect. As we stated in the Introduction, we are not the first to infer the effects of an intervention by examining the consequences of its discontinuation, but it is worth discussing the implications of this design for our research question. If practice exercises benefit students mainly by teaching them general (e.g., problem solving) skills, and the relationship between such benefits and the number of exercises attempted is concave (i.e., there are diminishing returns), our control-treatment comparisons will not capture such effects (because practice exercises will have impacted both experimental groups before our study), and we will under-estimate the effects of practice. Yet, if such exercises improve math achievement by exposing students to problems on specific domains (e.g., additions of two-digit fractions), allowing them to try different strategies, and observing which strategies work best (i.e., through content-specific trial and error), our impact estimates will capture such effects (because students were covering different topics, and thus completing different exercises, before and after the evaluation period).

Evidence from developmental psychology suggests that the type of transfer of knowledge and skills that would lead us to under-estimate the effect of independent practice is quite rare (see National Research Council 2000). While scholars in this field initially believed that practice with difficult tasks contributed to the development of general skills of learning and attention

3. Unfortunately, we do not have data on learning exercises; only on practice exercises (for treatment students).

4. We cannot calculate the share of units with practice exercises from our data, but we can calculate the percentage of all days in which students interacted with the CAL platform in which they also completed practice exercises. The average treatment student saw at least one practice exercise in 75% of all days in which he/she used the platform. Put differently, three of each four days that a treatment student used the software involved practice exercises.

5. We do not know the date in which each school started using the software, but assuming that all students before the start of evaluation attempted exercises at the same rate as treatment students during the evaluation period, we estimate that the former had been exposed to the software for 92 weeks (i.e., roughly two years).

(Woodworth and Thorndike 1901), experimental research discovered that, even when individuals seem to exhibit transfer, they are often relying on background knowledge that is of little use to solve similar tasks (e.g., Ericsson, Chase, and Faloon 1980). Subsequent studies have demonstrated that transfer depends not only on the degree to which tasks share common elements (Klausmeier 1985; Thorndike 1913), but also on the characteristics of learners (Singley and Anderson 1989), such as their mastery of the original subject (e.g., Klahr and Carver 1988; Littlefield et al. 1988), the extent to which they have learned it with understanding (instead of memorizing facts, e.g., Hendrickson and Schroeder 1941; Judd 1908), the time devoted to learning it (to develop pattern-recognition skills, e.g., Chase and Simon 1973; Ericsson, Krampe, and Tesch-Römer 1993; Simon and Chase 1973), and their motivation to learn (White 1959). Transfer is also affected by the context in which the original task was learned (e.g., Carraher, Carraher, and Schliemann 1985; Nunes et al. 1993) and the level of abstraction at which it was presented (e.g., Singley and Anderson 1989; Spiro et al. 1991). In fact, transfer may even be negative (i.e., experience with one set of events may actually hurt performance on related tasks) by leading individuals to adopt less efficient problem-solving strategies (Luchins and Luchins 1970). Therefore, our experimental design seems unlikely to under-estimate the effects of practice.

Given that we administered the baseline instruments on slightly different dates (see the Data section), some control students had access to the practice exercises after the baseline. Their exposure to this feature, however, was minimal: the mean student only spent 3.4 minutes attempting practice exercises during this period, and the median student 1.7 minutes. Thus, it is extremely unlikely that this brief exposure made a meaningful difference in their achievement.

2 Data

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

2.1 Student achievement

We administered student assessments of math learning at baseline (before the intervention) and endline (six months after the start of the intervention).[6] 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 three cognitive domains (knowing, applying, and reasoning). The distribution of items across content and cognitive domains was based on the assessment framework of the 2019 Trends in International Math and Science Study (TIMSS) for grade 4 (Mullis and Martin 2017).

Each test had 35 multiple-choice items. We drew on items from international assessments (e.g., TIMSS, PISA, Young Lives), domestic assessments (e.g., Quality Education Study, Student Learning Survey), and previous impact evaluations in India (e.g., the Andhra Pradesh Randomized Studies in Education or APRESt). We included items from a wide range of difficulty to reduce the possibility of "floor" effects (i.e., students not answering any questions correctly) and "ceiling" effects (i.e., students answering all questions correctly). At both baseline and endline, we administered one assessment for students in grades 4 and 5 and another one for students in grades 6 and 7, and we created four versions of each assessment to prevent cheating.[7]

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

2.2 Students' time on CAL platform and responses to exercises

We also obtained access to data on students' time interacting with the CAL platform and with the learning and practice exercises, as well as on whether their responses to these exercises were correct, to verify that the trial and the intervention were implemented as intended and to investigate the dose-response relationship between practice and achievement. Each student was assigned a unique login and password, which allowed us to track his/her usage and responses during the study. Using each student's Internet Protocol (IP) address and time of login, we could also determine whether he/she used the software in school or at home.

6. Different schools conducted the baseline assessment and started using the software on slightly different dates (see Table A3). However, throughout this paper, we limit our analysis to a common six-month period, starting on September 11, 2017 and ending on March 11, 2018.

7. The baseline tests can be accessed at: https://bit.ly/2MLWKqL (grades 4-5), https://bit.ly/3dSusXw (grades 6-7). The endline tests can be accessed at: https://bit.ly/2UtCSx6 (grades 4-5), https://bit.ly/30v0jcU (grades 6-7).

3 Empirical strategy

We estimate the effect of the offer of practice exercises (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 g and school s at time t (endline), $r(gs)$ is the randomization stratum of grade 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 interact the treatment dummy with students' grade, sex, and baseline achievement (continuous or by within-grade quartile) to understand whether the intervention is more helpful for some sub-groups of students.

We do not adjust our standard errors to account for classroom-level correlations in student outcomes because a recent study has demonstrated that such adjustment is not warranted when random assignment is conducted at the individual level (Abadie et al. 2017). Some have argued that this adjustment should be performed when randomization strata include very few observations (Chaisemartin and Ramirez-Cuellar 2020). Our strata are relatively large, but we demonstrate our estimates are robust to such adjustment in Table A9.

We compute family-wise error rate-adjusted p values in our estimation of heterogeneous ITT effects. Specifically, we follow List, Shaikh, and Xu (2019) and we account for pre-registered sub-group analyses by sex, enrolled grade, and initial math achievement.

4 Results

4.1 Implementation fidelity

The intervention was implemented largely as intended. First, virtually all students across both experimental groups (3,999 out of 4,001 students or 99.9%) logged in at least once to the CAL platform during the evaluation. In fact, the average student interacted with the software for 952 minutes during the six months of the study (i.e., about 38 minutes per week, see Figure 2), mostly at home (i.e., instead of at school, see Figure A2). Usage varied across schools typically between about 500 and 1,000 minutes; only one school logged more than 2,500 minutes (Figure A3, panel A). Variation in take-up is not uncommon, and it may have been due to differences in school infrastructure (e.g., availability of computers), teacher buy-in, and/or alignment between the software and the school's math curriculum, among other factors.

Second, all treatment students attempted at least one practice exercise during the study. The average treatment student spent 76 minutes attempting practice exercises during the study period (i.e., about 3 minutes per week, see Figure 3). Interaction with the practice exercises also differed across schools, following a similar pattern as software usage (Figure A3, panel B).

Lastly, the practice exercises that treatment students attempted covered 18 topics (e.g., "geometry"), 59 subtopics (e.g., "triangles and triangle properties"), and 151 units within those topics (e.g., "classifying triangles based on sides and angles"). The three topics in which students attempted the most amount of practice exercises were measurement (14% of the total number of exercises), fractions (10%), and number theory (9.1%).

4.2 Average effects on math achievement

The offer of the intervention had a precisely estimated null effect (of 0.01 standard deviations or SDs) on the math achievement of the average student, regardless of whether we account for students' baseline performance or not (Table 2). In fact, based on the 95% confidence interval, we can rule out effects below -0.04 SDs and above 0.06 SDs. This null average effect is consistent across content domains, with point estimates ranging from 0.002 to 0.01, and across cognitive domains, with point estimates ranging from 0.002 to 0.01 (Table 3).[8] It is also consistent across "repeated" items across baseline and endline and "non-repeated" items (introduced in the endline; Table A5).

4.3 Heterogeneous effects on math achievement

We explored whether the effect of the intervention differed across the only three student characteristics recorded in our data—sex, enrolled grade, and initial achievement—as we had specified in our pre-analysis plan. We found that the intervention had a moderate-to-large positive effect of 0.14 SDs for students with initially low math achievement. First, we do so graphically. We plot the effects of the intervention for each within-grade quartile of baseline math achievement (Figure 4). Then, we do so analytically. We examine this heterogeneity in two ways: first, by interacting the treatment indicator variable with each student's (continuous) baseline score and then, by interacting that indicator with indicators for each student's within-grade quartile of baseline achievement (Table 4). Even after adjusting the p values for multiple-hypothesis testing, our estimate of the causal effect of practice exercises on initially low-achieving students remained statistically significant at the 1% level ($p = 0.001$, not in table).

We explored, but ultimately rejected, the possibility that the effects of the intervention on initially low-achieving students could be explained by impacts on English-language proficiency. Specifically,

8. In the reasoning cognitive domain, we find a marginally statistically significant effect, which is likely due to the number of hypothesis tests that we are running (see Table 3).

given that instruction in private-unaided schools in India is delivered in English (see Singh and Sarkar 2015), and our baseline and endline assessments were also in English (rather than in Indian vernacular languages), it seems at least possible that practice exercises may have improved the math achievement of low performers by teaching them English (math) words. We examined this possibility in three ways. First, we calculated students' proportion-correct scores on the endline assessment for items that did not require them to read words (beyond a short prompt, such as "solve") and for items without "math vocabulary" words (e.g., "rectangle"), and we estimated the impact of the intervention on each of these separately. Then, we used a confirmatory, two-dimensional item-response theory (IRT) model to scale students' endline scores by allowing items that did not require literacy and those that did to load onto separate correlated factors, and we estimated the effect of the intervention on the latter. Next, we used a similar IRT model in which all items loaded onto the first factor but those requiring literacy may also load to a correlated factor, and we estimated the effect of the intervention on the former. If initially low achievers only improved because they were "taught" English through practice exercises, they should not have improved (or they should have improved less) on the aforementioned six outcome variables. However, this was not the case. In fact, impacts on these outcomes were indistinguishably close to our main estimates in Tables 2 and 4 (see Table A6).

We did not, however, find compelling evidence of heterogeneous effects by other factors (see Table A7). Female students performed slightly below male students, but the difference was not statistically significant, nor was the interaction between the treatment and female indicator. The effect of the intervention did not vary by the grade in which students were enrolled either. While we observed some variation in the effect of the intervention across schools, we were under-powered to detect statistically significant differences (see Figure A4).

4.4 Average effects on interaction with CAL software

Finally, we considered whether practice exercises impacted the extent to which students interacted with the CAL software. First, given that treatment students were required to attempt practice exercises at the end of each unit but control students were not (see Intervention section), it is possible that the intervention led the former to complete fewer units in the CAL platform than the latter. This would have been problematic because, while we expected practice exercises to *positively* impact the math achievement of treatment students, we would expect completing fewer units to *negatively* impact their achievement, so the overall effect of the intervention would have confounded these conflicting influences. We addressed this possibility in the first column of Table A8, where we estimated the effect of the intervention on the number of sessions completed in the CAL platform. Treatment students spent only about 1% more of sessions than control students, and the difference between the two was statistically insignificant. Second, given that we evaluated the effect of practice

exercises by deactivating them for control students during the study period (see Intervention section), it is possible that they compensated for this change by spending more time on the platform than their treatment peers. This would have also been problematic because it would have led us to confound the effect of exposure to practice exercises with the effect of spending time on the CAL platform. We addressed this possibility in the second column of Table A8, where we estimated the effect of the intervention on the *total time* spent on the platform. Treatment students spent roughly 2.3% more minutes than control students, but again, the difference was not statistically significant. In the third column of the table, we also considered the possibility that the intervention affected the time that control students spent on the platform during *each* session by estimating the effect of the intervention on the total time spent on the platform, holding the number of sessions constant. Per session, treatment students spent approximately 1.5% more minutes on the platform, but the difference was only marginally statistically significant. In short, we did not find much evidence that the intervention impacted the extent to which students interacted with the CAL software, suggesting that we were not confounding the effects of practice with those of other factors.

5 Conclusion

This paper presents one of the first studies that is sufficiently powered to simultaneously rule out meaningful effects of independent practice for the average and detect non-trivial positive effects for lower-performing students. After only six months of an average of three minutes of practice per day, we find that the lowest-performing students attending private schools that cater to relatively well-off families outperformed their control peers by 0.14 SDs.[9] However, this extra practice had no effect on their average-performing counterparts. These results are robust across sub-group analyses and different schools and cities.

Our study makes an important contribution to three different but related literatures. First and foremost, it identifies an approach to address heterogeneity in students' preparation for schooling, a frontier challenge in developing countries (see Ganimian and Murnane 2016; Glewwe and Muralidharan 2016). This approach demands far less time from students and teachers than remedial interventions with similar effects (e.g., A. V. Banerjee et al. 2007b). It can be pursued (mostly) after school hours, without requiring that teachers divert from the curriculum, or that they take time away from core subjects to take students to the computer lab—two factors that have frustrated efforts to scale-up similarly effective interventions (see, e.g., Muralidharan and Singh 2019; A. Banerjee et al. 2017). And, conditional on the availability of the requisite hardware and software required for the CAL platform in which it is embedded (in this case, Mindspark) it does

9. We use the term "extra practice" to refer to the practice exercises that treatment students, but not control students, were expected to complete during the study period (this was the intervention). As we explain in the Results section, however, control and treatment students spent the same amount of time with the CAL software.

not require any additional setup costs or training. Given the wide reach of the Mindspark software in India and abroad, and the current need to educate students while they are out of school due to the coronavirus, our results suggest that we could be helping a lot more students catch up with their peers by encouraging independent practice.

Second, our study also contributes to the rapidly evolving body of research on private schools in India and in developing countries more generally. While our study is by no means representative of all private-unaided schools, it sheds light on a sector that has received considerably less attention than it warrants in light of its domestic and global share of student enrollment in elementary education. Our study identifies an approach through which at least unaided private schools could improve learning among low performers. Given the low costs of the underlying platform (see Muralidharan, Singh, and Ganimian 2019), we are optimistic that a similar approach could be adopted in aided private schools, which make up the bulk of the sector in South Asia (see Andrabi et al. 2007).

Finally, our study offers an important demonstration of how to leverage the growing penetration of educational software products to run rapid-cycle randomized evaluations that shed light on the merits of intuitively appealing yet largely untested educational strategies while reducing the time and cost required for independent data collection. Perhaps more importantly, it does so in a way that allows researchers to closely monitor students' interaction with the intervention being tested (in this case, independent practice) and to estimate its effect precisely, not only for the average student, but also for relevant sub-groups. We see this as a crucial contribution to research on education technology, given that many interventions that have been evaluated in this space have yielded disappointing results and would benefit from feedback on their effectiveness (Ganimian, Hess, and Vegas 2020).

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* Technical report w24003. Cambridge, MA: National Bureau of Economic Research.

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

Argys, Laura M., Andrew I. Friedson, M. Melinda Pitts, and D. Sebastian Tello-Trillo. 2020. "Losing public health insurance: TennCare reform and personal financial distress." *Journal of Public Economics* 187 (July): 104202.

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

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31, no. 4 (November): 73–102.

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

———. 2007b. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122 (3): 1235–1264.

Bettinger, Eric, Robert W Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov. 2020. *Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function.* Working Paper 26967. Cambridge, MA: National Bureau of Economic Research, April.

Cardelle-Elawar, Maria. 1990. "Effects of Feedback Tailored to Bilingual Students' Mathematics Needs on Verbal Problem Solving." *The Elementary School Journal* 91, no. 2 (November): 165–175.

———. 1995. "Effects of metacognitive instruction on low achievers in mathematics problems." *Teaching and Teacher Education* 11, no. 1 (January): 81–95.

Carraher, Terezinha Nunes, David William Carraher, and Analúcia Dias Schliemann. 1985. "Mathematics in the streets and in schools." *British Journal of Developmental Psychology* 3 (1): 21–29.

Central Square Foundation. 2020. *State of the Sector Report on Private Schools in India.* Report. New Delhi: Central Square Foundation.

Chaisemartin, Clément de, and Jaime Ramirez-Cuellar. 2020. *At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?* Working Paper 27609. Cambridge, MA: National Bureau of Economic Research, August.

Chase, William G., and Herbert A. Simon. 1973. "Perception in chess." *Cognitive Psychology* 4 (1): 55–81.

Chyn, Eric. 2018. "Moved to Opportunity: The Long-Run Effects of Public Housing Demolition on Children." *American Economic Review* 108, no. 10 (October): 3028–3056.

Dynarski, Susan M. 2003. "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion." *American Economic Review* 93, no. 1 (March): 279–288.

Ericsson, K. Anders, William G. Chase, and Steve Faloon. 1980. "Acquisition of a Memory Skill." *Science* 208, no. 4448 (June): 1181–1182.

Ericsson, K. Anders, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. "The role of deliberate practice in the acquisition of expert performance." *Psychological Review* 100, no. 3 (July): 363–406.

Fishman, Ram, Stephen C Smith, Vida Bobić, and Munshi Sulaiman. 2017. *How Sustainable Are Benefits from Extension for Smallholder Farmers? Evidence from a Randomized Phase-Out of the BRAC Program in Uganda.* Working Paper. Bonn, Germany: IZA Institute of Labor Economics.

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

Ganimian, Alejandro J., and Richard J. Murnane. 2016. "Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations." *Review of Educational Research* 86 (3): 719–755.

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

Hendrickson, G., and W. H. Schroeder. 1941. "Transfer of training in learning to hit a submerged target." *Journal of Educational Psychology* 32 (3): 205–213.

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

Hohn, Robert L., and Bruce Frey. 2002. "Heuristic Training and Performance in Elementary Mathematical Problem Solving." *The Journal of Educational Research* 95, no. 6 (July): 374–380.

Jitendra, A. K., C. C. Griffin, K. McGoey, M. C. Gardill, P. Bhat, and T. Riley. 1998. "Effects of mathematical word problem solving by students at risk or with mild disabilities." *The Journal of Educational Research* 91 (6): 345–355.

Jitendra, Asha K., Jon R. Star, Michael Rodriguez, Mary Lindell, and Fumio Someki. 2011. "Improving students' proportional thinking using schema-based instruction." *Learning and Instruction* 21, no. 6 (December): 731–745.

Judd, C. H. 1908. "The relation of special training and general intelligence." *Educational Review* 36:28–42.

King, Alison. 1991. "Effects of training in strategic questioning on children's problem-solving performance." *Journal of Educational Psychology* 83 (3): 307–317.

Klahr, David, and Sharon McCoy Carver. 1988. "Cognitive objectives in a LOGO debugging curriculum: Instruction, learning, and transfer." *Cognitive Psychology* 20, no. 3 (July): 362–404.

Klausmeier, Herbert J. 1985. *Educational psychology.* 5. ed. New York: Harper & Row.

Kolen, Michael J, and Robert L Brennan. 2004. *Test Equating, Scaling, and Linking.* 3rd. New York, NY: Springer.

Kramarski, Bracha, and Zemira R. Mevarech. 2003. "Enhancing Mathematical Reasoning in the Classroom: The Effects of Cooperative Learning and Metacognitive Training." *American Educational Research Journal* 40, no. 1 (January): 281–310.

Lai, Fang, Renfu Luo, Linxiu Zhang, Xinzhe Huang, and Scott Rozelle. 2015. "Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing." *Economics of Education Review* 47 (August): 34–48.

Lai, Fang, Linxiu Zhang, Xiao Hu, Qinghe Qu, Yaojiang Shi, Yajie Qiao, Matthew Boswell, and Scott Rozelle. 2013. "Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in Shaanxi." *Journal of Development Effectiveness* 52 (2): 208–231.

Lai, Fang, Linxiu Zhang, Qinghe Qu, Xiao Hu, Yaojiang Shi, Matthew Boswell, and Scott Rozelle. 2016. *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai, China.* Working Paper 237. Stanford, CA: Stanford University, Freeman Spogli Institute for International Studies, August.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76, no. 3 (July): 1071–1102.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple hypothesis testing in experimental economics." *Experimental Economics* 22, no. 4 (December): 773–793.

Littlefield, Joan, Victor R. Delclos, Sharon Lever, Keith N. Clayton, John D. Bransford, and Jeffery J. Franks. 1988. "Learning LOGO: Method of teaching, transfer of general skills, and attitudes toward school and computers." In *Teaching and learning computer programming: Multiple research perspectives,* edited by Richard E. Mayer, 111–135. Lawrence Erlbaum Associates, Inc.

Luchins, Abraham S., and Edith H. Luchins. 1970. *Wertheimer's Seminars Revisited: Problem Solving and Thinking, Volume I.* 1st edition. Albany, NY: State University of New York at Albany, January.

Mevarech, Zemira R., and Bracha Kramarski. 2003. "The effects of metacognitive training versus worked-out examples on students' mathematical reasoning." *British Journal of Educational Psychology* 73 (4): 449–471.

Mo, D., L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell, and S. Rozelle. 2015. "Persistence of learning gains from computer assisted learning: Experimental evidence from China." *Journal of Computer Assisted Learning* 31 (6): 562–581.

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

Muralidharan, Karthik, and Michael Kremer. 2008. "Public-Private Schools in Rural India." In *School Choice International: Exploring Public-Private Partnerships,* edited by Rajashri Chakrabarti and Paul E. Peterson, 91–110. Cambridge, MA: MIT Press.

Muralidharan, Karthik, and Abhijeet Singh. 2019. "Improving Schooling Productivity through Computer-Aided Personalization: Experimental Evidence from Rajasthan." Washington, D.C.: RISE Annual Conference 2019, June.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109, no. 4 (April): 1426–1460.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy* 119 (1): 39–77.

———. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *The Quarterly Journal of Economics* 130, no. 3 (August): 1011–1066.

Nakajima, Nozomi. 2020. "Long-run effects of short-term grants in early childhood education." Cambridge, MA, December.

National Research Council. 2000. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition.* Washington, D.C.: National Academies Press, August.

NCERT. 2017. *National Achievement Survey (NAS) dashboard: 2017.*

NCES. 2019. *Digest of education statistics: 2019.*

NIEPA. 2018a. *Elementary Education in India: Where Do We Stand? District Report Cards 2016-17.* Vol. 1. New Delhi, India: National Institute of Educational Planning / Administration.

———. 2018b. *Elementary Education in India: Where Do We Stand? District Report Cards 2016-17.* Vol. 2. New Delhi, India: National Institute of Educational Planning / Administration.

———. 2019. *U-DISE Flash Statistics 2017-18.* New Delhi, India: National Institute of Educational Planning / Administration.

Nunes, Terezinha, Terezinha Nunes Carraher, Analucia Dias Schliemann, and David William Carraher. 1993. *Street Mathematics and School Mathematics.* Cambridge University Press, April.

Palinscar, Aannemarie Sullivan, and Ann L. Brown. 1984. "Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities." *Cognition and Instruction* 1, no. 2 (March): 117–175.

Pitchford, Nicola J. 2015. "Development of early mathematical skills with a tablet intervention: a randomized control trial in Malawi." *Frontiers in Psychology* 6 (April).

Scardamalia, Marlene, Carl Bereiter, and Rosanne Steinbach. 1984. "Teachability of Reflective Processes in Written Composition." *Cognitive Science* 8, no. 2 (April): 173–190.

Simon, Herbert A., and William G. Chase. 1973. "Skill in chess." *American Scientist* 61 (4): 394–403.

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

Singh, Renu, and Sudipa Sarkar. 2015. "Does teaching quality matter? Students learning outcome related to teaching quality in public and private primary schools in India." *International Journal of Educational Development* 41 (March): 153–163.

Singley, Mark K., and John Robert Anderson. 1989. *The Transfer of Cognitive Skill.* Harvard University Press.

Spiro, Rand J., Paul J. Feltovich, Paul L. Feltovich, Michael J. Jacobson, and Richard L. Coulson. 1991. "Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains." *Educational Technology* 31 (5): 24–33.

Thorndike, Edward L. 1913. *Educational psychology.* New York: Teachers College, Columbia University.

UIS-UNESCO. 2019. *UIS statistics.*

Verschaffel, Lieven, Erik De Corte, Sabien Lasure, Griet Van Vaerenbergh, Hedwig Bogaerts, and Elie Ratinckx. 1999. "Learning to Solve Mathematical Application Problems: A Design Experiment With Fifth Graders." *Mathematical Thinking and Learning* 1, no. 3 (September): 195–229.

White, Robert W. 1959. "Motivation reconsidered: The concept of competence." *Psychological Review* 66 (5): 297–333.

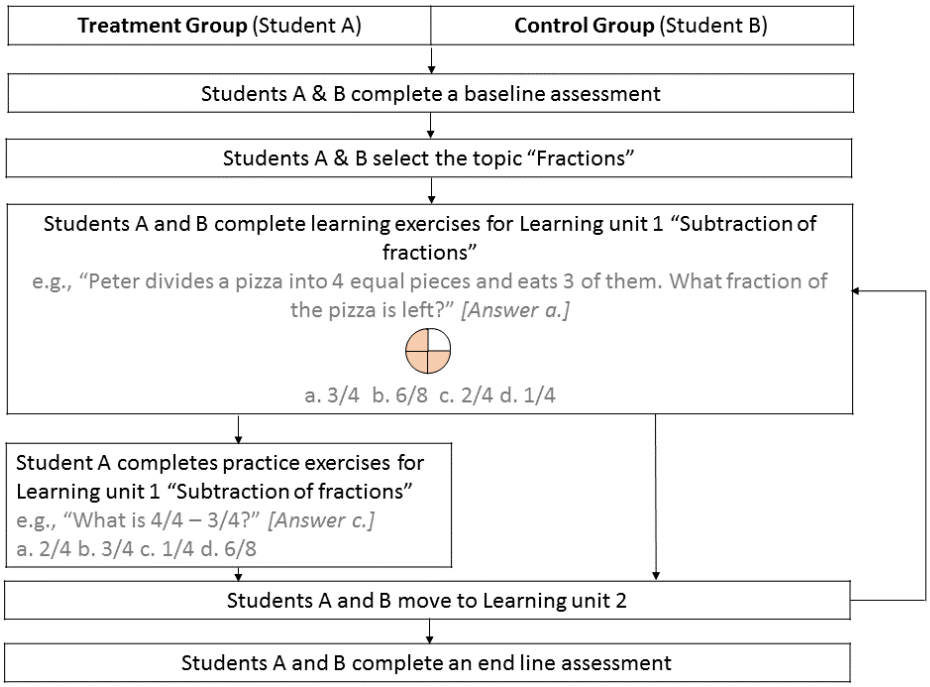
Woodward, John, Sybilla Beckmann, Mark Driscoll, Megan Franke, Patricia Herzig, Asha Jitendra, Kenneth R. Koedinger, and Philip Ogbuehi. 2012. *Improving mathematical problem solving in grades 4 through 8: A practice guide.* Practice Guide NCEE 2012-4055. Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Woodworth, R. S., and E. L. Thorndike. 1901. "The influence of improvement in one mental function upon the efficiency of other functions. (I)." *Psychological Review* 8 (3): 247–261.

6 Figures and Tables

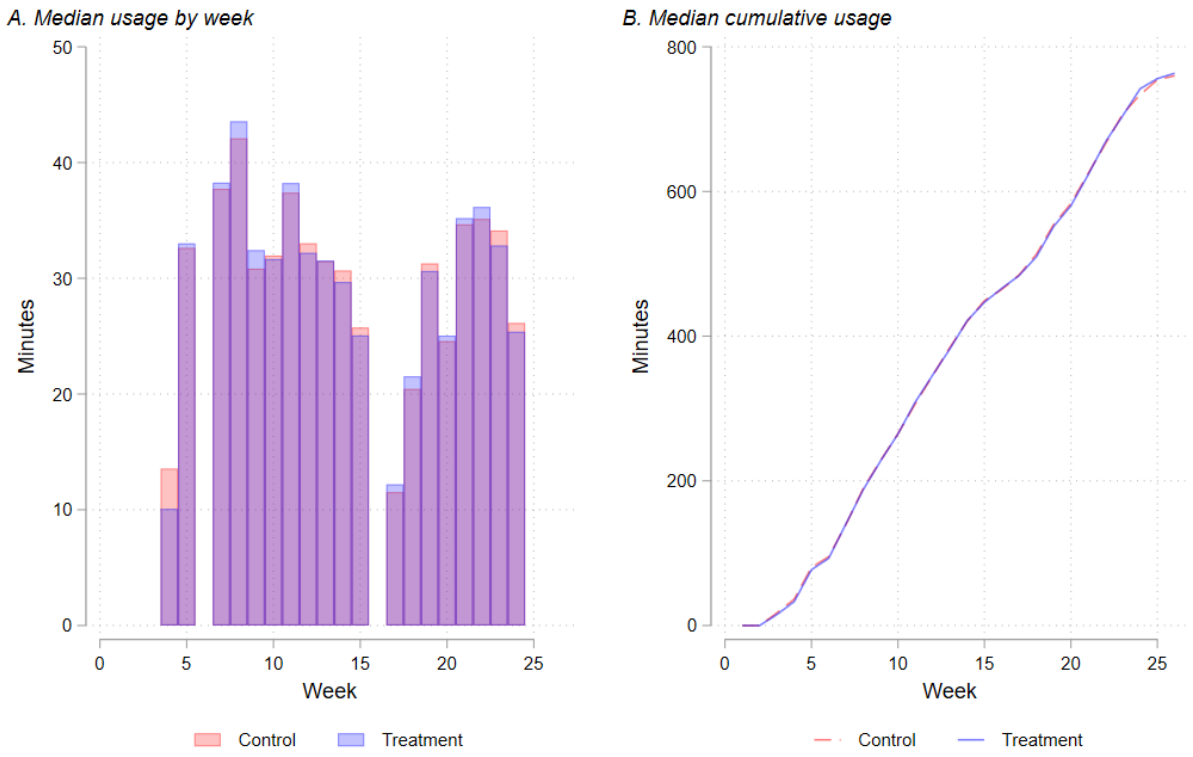
Figures

Figure 1: *Differences between the sequence of exercises completed by students in the study*



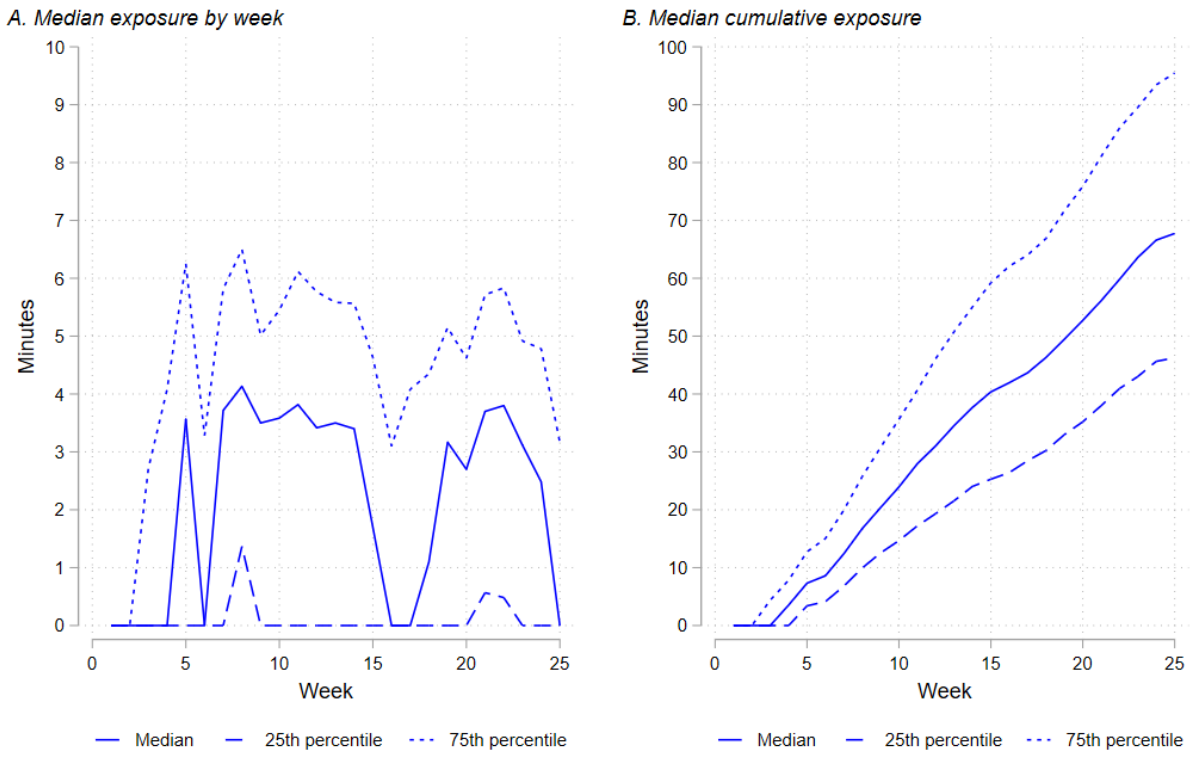
Note: This figure describes the sequence of exercises completed by students in both experimental groups in the study.

Figure 2: *Weekly and cumulative time spent on the CAL platform during the study*



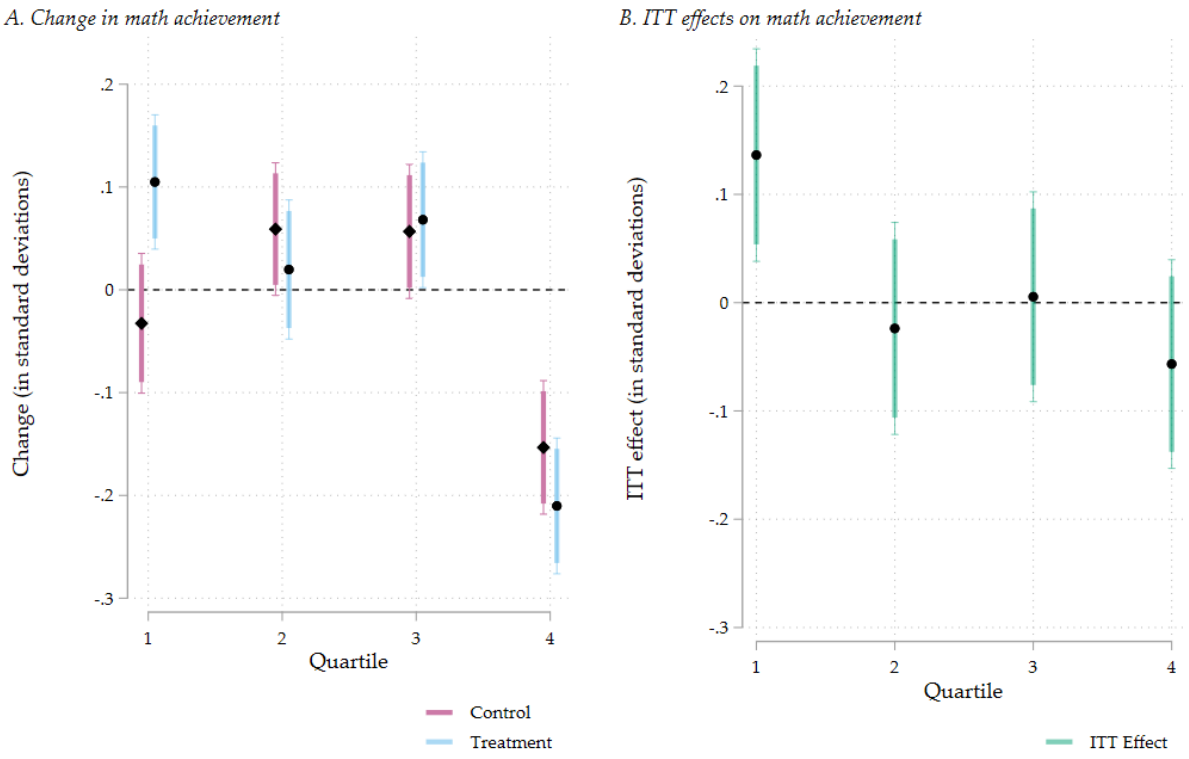
Notes: (1) This figure shows the weekly (panel A) and cumulative (panel B) usage of the CAL platform for the median student, by experimental group. (2) This figure includes all students observed at baseline and endline, regardless of whether they used the software (99.9% of students did). (3) Usage is binned by weeks elapsed since the start of the study (on September 11, 2017).

Figure 3: *Weekly and cumulative time spent on practice exercises during the study among treatment students*



Notes: (1) This figure shows the weekly (panel A) and cumulative (panel B) time spent on practice exercises platform for three groups of treatment students: the median student (i.e., 50th percentile), the 25th percentile, and the 75th percentile, to give a sense of variation in usage in the sample. (2) This figure includes all treatment students observed at baseline and endline, regardless of whether they used the software (99.9% of students did). (3) Usage is binned by weeks elapsed since the start of the study (on September 11, 2017).

Figure 4: *Heterogeneous ITT effects on math achievement at endline, by quartile of baseline performance*



Notes: ((1) This figure shows heterogeneity in the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after six months), by within-grade quartile of baseline performance. (2) Both panels account for randomization-strata fixed effects. (3) Bars and whiskers show 90-percent and 95-percent confidence intervals, respectively.

Tables

Table 1: *Balancing checks between experimental groups*

	(1) Control	(2) Treatment	(3) Difference
A. Grade-wise distribution (full sample)			
Grade 4	0.24	0.25	-0.00
	[0.43]	[0.43]	
Grade 5	0.29	0.28	0.01
	[0.45]	[0.45]	
Grade 6	0.25	0.26	-0.01
	[0.43]	[0.44]	
Grade 7	0.22	0.22	0.00
	[0.41]	[0.41]	
B. Balance tests (full sample)			
Baseline score	0.00	-0.03	0.03
	[1.00]	[1.01]	(0.02)
Female	0.53	0.51	0.02
	[0.50]	[0.50]	(0.01)
N (students)	2234	2227	
C. Balance tests (non-attritors)			
Baseline score	0.02	-0.01	0.04
	[1.00]	[1.00]	(0.02)
Female	0.53	0.52	0.01
	[0.50]	[0.50]	(0.01)
N (students)	1984	2017	

Notes: (1) 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 compares grade enrollment. It does not perform significance tests because, due to the stratification strategy, grade enrollment is comparable across 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 (90% of the total). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: *ITT effect of practice exercises on math achievement at endline*

	Math (IRT-scaled) score	
	(1)	(2)
Treatment	0.014	0.014
	(0.027)	(0.025)
Baseline score		0.645***
		(0.025)
N (students)	4001	4001
R-squared	0.512	0.588

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after six months). Column 1 shows the simple difference in means (including randomization-strata fixed effects); column 2 moreover controls for baseline score. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

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

	A. All students					
	(1) Numbers	(2) Geometry	(3) Data	(4) Knowing	(5) Applying	(6) Reasoning
Treatment	0.002	0.008	0.002	0.004	0.002	0.011*
	(0.004)	(0.005)	(0.006)	(0.005)	(0.004)	(0.006)
Baseline score	0.077***	0.117***	0.114***	0.065***	0.102***	0.132***
	(0.004)	(0.005)	(0.006)	(0.005)	(0.004)	(0.006)
N (students)	4001	4001	4001	4001	4001	4001
R-squared	0.513	0.495	0.405	0.475	0.533	0.501
	B. Low-performing students					
	(1) Numbers	(2) Geometry	(3) Data	(4) Knowing	(5) Applying	(6) Reasoning
Treatment	0.020**	0.051***	0.013	0.027***	0.022**	0.042***
	(0.009)	(0.011)	(0.012)	(0.009)	(0.009)	(0.013)
Baseline score	0.088***	0.125***	0.115***	0.076***	0.110***	0.141***
	(0.007)	(0.009)	(0.010)	(0.008)	(0.007)	(0.011)
N (students)	4001	4001	4001	4001	4001	4001
R-squared	0.517	0.503	0.408	0.478	0.538	0.506

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in each content (columns 1-3) and cognitive (columns 4-6) domain at endline (after six months). 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. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

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

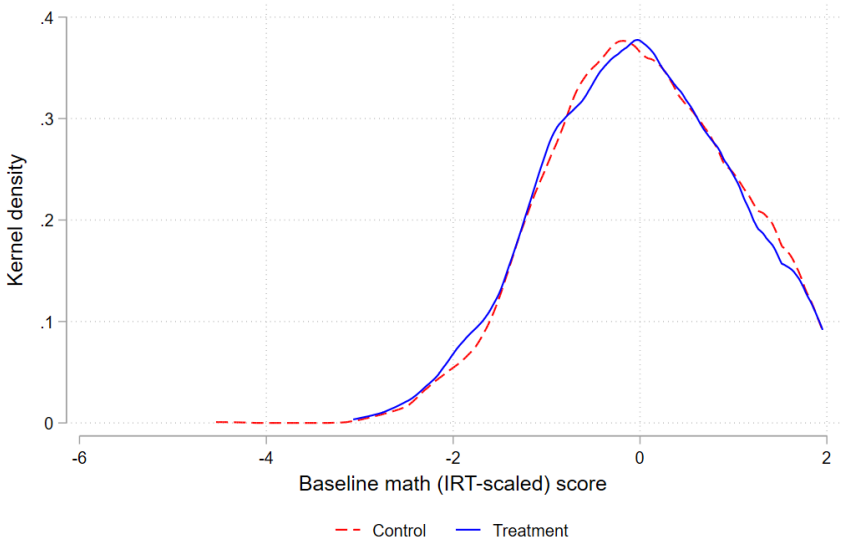
	Math (IRT-scaled) score	
	(1)	(2)
Treatment	0.014	0.136***
	(0.025)	(0.050)
Baseline score	0.672***	0.647***
	(0.028)	(0.042)
Treatment X Baseline	-0.054**	
	(0.025)	
Quartile 2		0.182***
		(0.065)
Quartile 3		0.198**
		(0.085)
Quartile 4		0.144
		(0.110)
Treatment X Quartile 2		-0.160**
		(0.071)
Treatment X Quartile 3		-0.131*
		(0.071)
Treatment X Quartile 4		-0.193***
		(0.070)
N (students)	4001	4001
R-squared	0.588	0.590

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after six months) by baseline performance, either as a continuous score (column 1) or as a set of quartile indicator variables (column 2). All estimations include randomization-strata fixed effects. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A Additional figures and tables

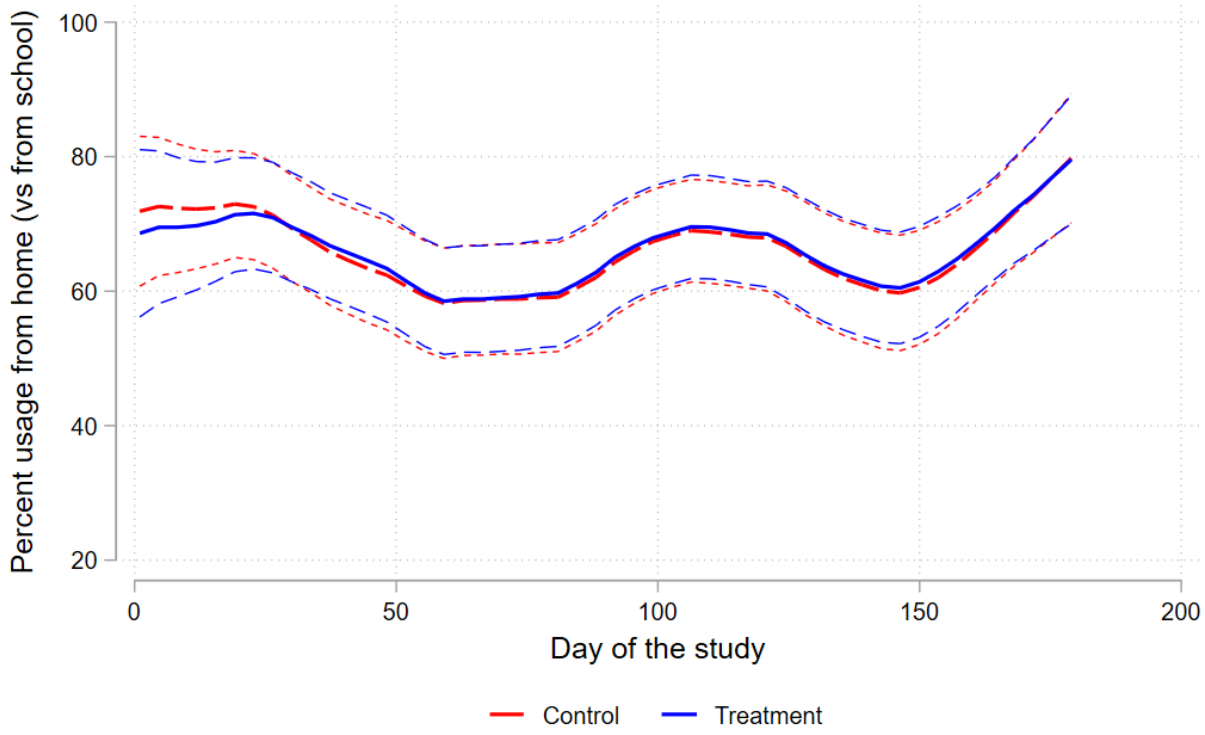
Additional figures

Figure A1: *Distribution of math (IRT-scaled) scores by experimental group at baseline*



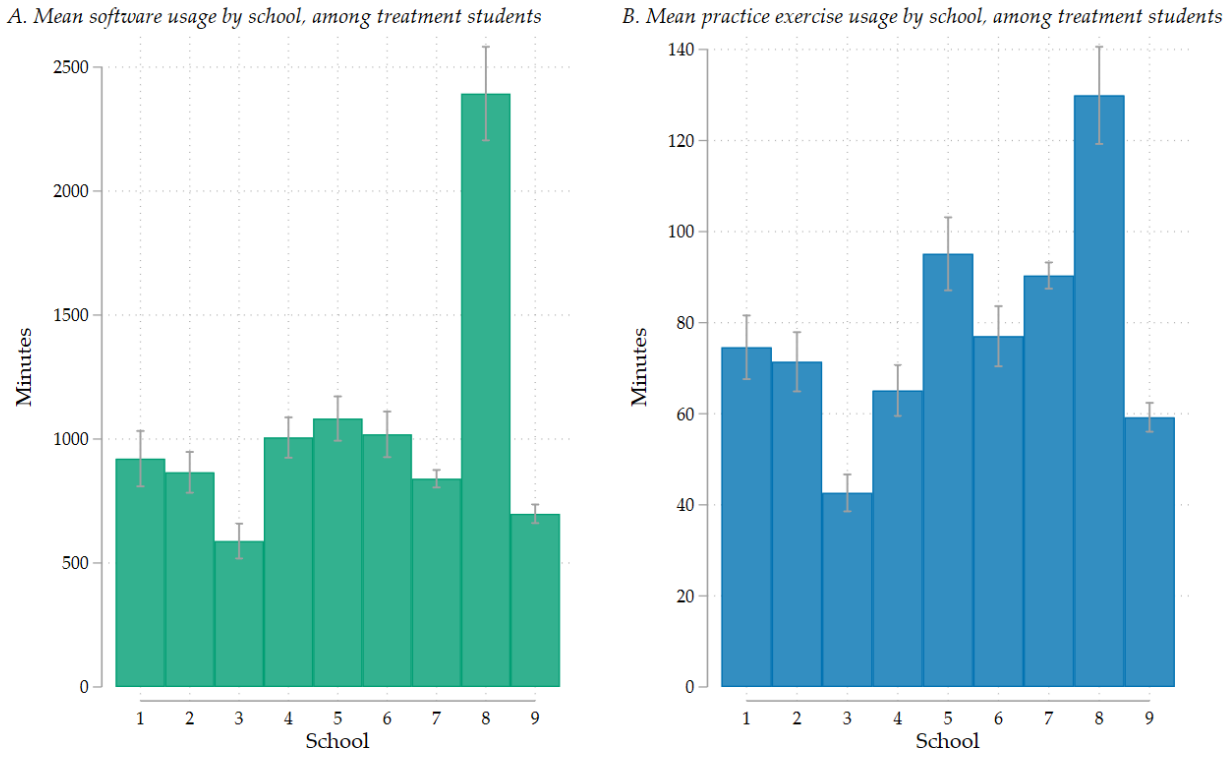
Notes: (1) This figure shows the distribution of scores in the baseline assessment of math for control and treatment students. (2) Scores were scaled using a two-parameter logistic Item Response Theory (IRT) model. (3) This figure includes all students present at baseline and endline.

Figure A2: *Percentage of time spent on the CAL platform at home (instead of at school)*



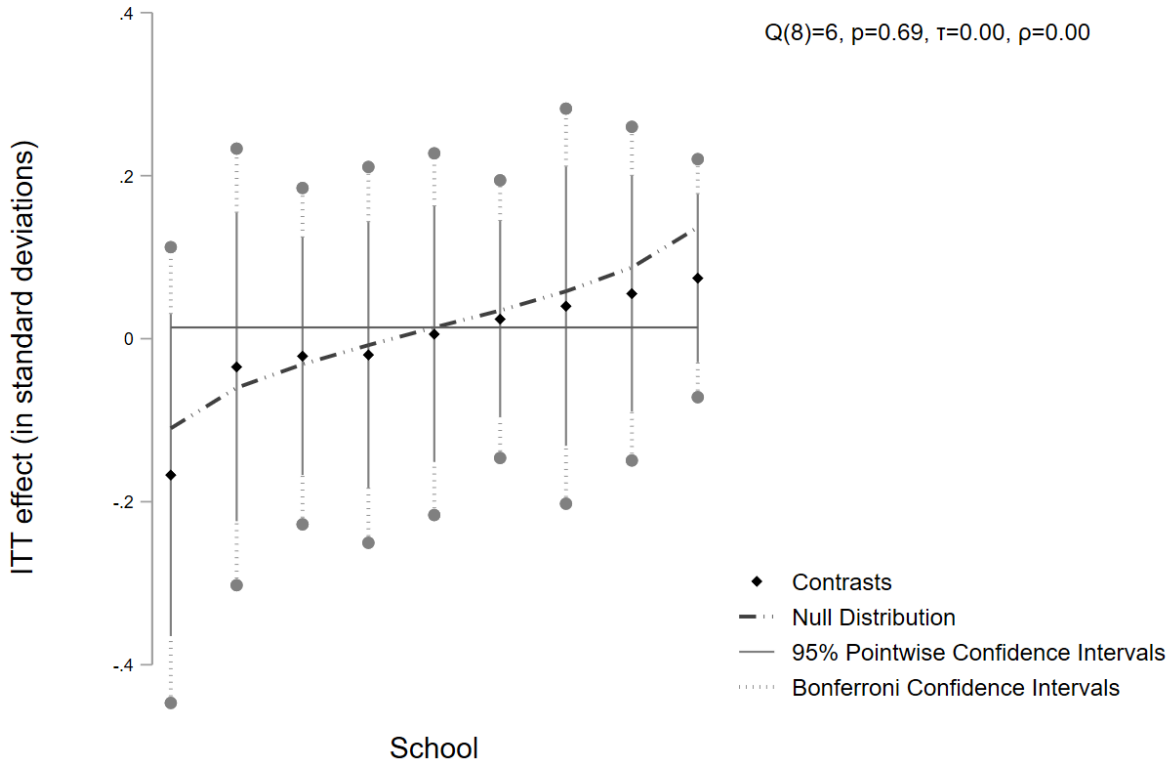
Notes: (1) This figure shows a local polynomial smooth plot with confidence intervals, for the number of minutes that the average student spent using the CAL software at home (rather than during school hours) for each day of the study. (2) 95-percent confidence intervals shown with dashed lines. (3) This figure includes all students who used the software on a given day.

Figure A3: *Take-up among treatment-group students, by school*



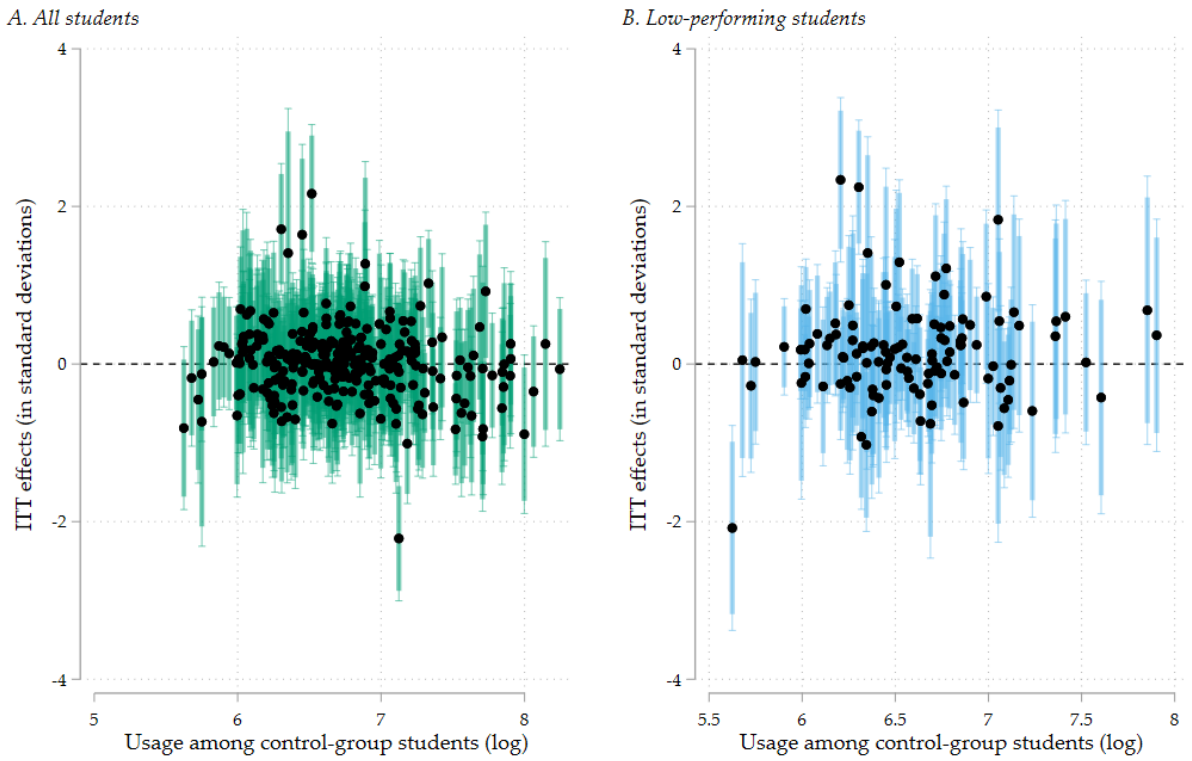
Notes: (1) By study school, this figure shows the average number of minutes treatment-group students spent on the software (panel A) and on practice exercises (panel B). (2) Standard errors shown with gray whiskers. (3) School numbers are in alphabetical order of locations, as follows. 1: Ahmedabad, Gujarat; 2: Faridabad, Haryana; 3: Ghaziabad, Uttar Pradesh; 4-5: Kolkata, West Bengal; 6-7: New Delhi, Delhi; 8: Rajkot, Gujarat; 9: Tiruchirappalli, Tamil Nadu.

Figure A4: *Heterogeneous ITT effects on math achievement at endline, by school*



Notes: (1) This figure provides a "caterpillar plot" of ITT effects by school (cf. Hippel and Bellows 2018). Each black dot refers to the point estimate for a given school. (2) 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). (3) The estimation controls for student baseline achievement and randomization-strata fixed effects.

Figure A5: *Dose-response relationship*



Notes: (1) This figure shows heterogeneity in the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after six months) by randomization stratum, for all students (panel A) and students in the bottom quartile of baseline achievement within their grade level (panel B). (2) Bars and whiskers show 90-percent and 95-percent confidence intervals, respectively.

Additional tables

Table A1: *Distribution of elementary-education student enrollment by school type in Indian states represented in this study, 2006-2007 and 2017-2018*

	(1)	(2)	(3)	(4)	(5)	(6)
	Share of enrollment in grades 1 to 8 (elementary education)					
	Government		Private unaided		Private aided	
State	2007-08	2017-18	2007-08	2017-18	2007-08	2017-18
Delhi	66%	53%	28%	44%	5.40%	3.40%
Gujarat	79%	62%	19%	36%	2.80%	2.10%
Haryana	74%	39%	23%	59%	3.20%	1.50%
Tamil Nadu	50%	37%	25%	46%	25%	17%
Uttar Pradesh	70%	46%	30%	42%	5.40%	5.60%
West Bengal	87%	86%	8.10%	11%	4.90%	0.30%

Sources: Central Square Foundation (2020).
Notes: (1) Percentages indicate the share of students enrolled in government, private-unaided, and private-aided schools for all states represented in our study.

Table A2: *Share of practice exercises below, at, or above grade level*

Enrolled grade level	Share of practice exercises...				
	...two or more grade levels behind	...one grade level behind	...at grade level	...one grade level above	...two or more grade levels above
Grade 4	0.001	0.070	0.422	0.330	0.177
Grade 5	0.019	0.065	0.407	0.315	0.194
Grade 6	0.010	0.066	0.881	0.042	0.000
Grade 7	0.010	0.281	0.632	0.078	0.000
All grades	0.010	0.113	0.579	0.198	0.099

Notes: (1) This table shows the share of practice exercises completed by treatment students during the study by the grade level in which students were enrolled and the grade level in which each exercise was categorized. Specifically, it shows the share of exercises one or two (or more) grade levels below, at grade level, or one or two (or more) grade levels above the enrolled grade of each student. (2) Practice exercises can be mapped to multiple levels. In this table, if an exercise includes at least the student's enrolled grade level, it is marked as at-level.

Table A3: *Cities, schools, assessment, and software-activation dates*

	(1) School	(2) Baseline date (2017)	(3) Activation date (2017)	(4) Endline date (2018)
Ahmedabad, Gujarat	1	25-26 Sep	6-Oct	3-5 Apr
Faridabad, Haryana	2	26-Sep	6-Oct	27-28 Mar
Ghaziabad, Uttar Pradesh	3	26-Sep	6-Oct	13-14 Apr
Kolkata, West Bengal	4	15-Sep	15-Sep	7-9 Mar
	5	14-Sep	17-Oct	5-7 Mar
New Delhi, Delhi	6	21-22 Sep	17-Oct	16-Apr
	7	25-Sep	6-Oct	9, 11, 12 Apr
Rajkot, Gujarat	8	20-Sep	22-Sep	15-Mar
Tiruchirappalli, Tamil Nadu	9	7-Oct	9-Oct	9, 11-13 Apr

Notes: (1) This table shows the list of sites, schools, assessment, and software activation dates for the study sample. (2) Software activation date refers to the date in which the practice exercises were made unavailable to control students. (3) Schools with multiple baseline and endline dates had multiple grades in the study, which differed on their test dates.

Table A4: *Balancing checks between experimental groups (among low-performing students only)*

	(1) Control Mean/SD	(2) Treatment Mean/SD	Difference (1)-(2)
A. Grade-wise distribution (full sample)			
Grade 4	0.25	0.24	0.01
	[0.43]	[0.43]	
Grade 5	0.29	0.28	0.01
	[0.45]	[0.45]	
Grade 6	0.25	0.26	–0.01
	[0.43]	[0.44]	
Grade 7	0.21	0.22	–0.02
	[0.41]	[0.42]	
B. Balance tests (full sample)			
Baseline score	–1.18	–1.15	–0.03
	[0.63]	[0.63]	(0.03)
Female	0.49	0.49	–0.00
	[0.50]	[0.50]	(0.03)
N (students)	533	584	
C. Balance tests (non-attritors)			
Baseline score	–1.16	–1.15	–0.01
	[0.62]	[0.62]	(0.03)
Female	0.50	0.50	–0.00
	[0.50]	[0.50]	(0.03)
N (students)	462	519	

Notes: (1) This table compares low-performing 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). "Low-performing" refers to students in a grade-level's bottom quartile, as per students' performance on the baseline assessment. Panel A compares grade enrollment. It does not perform significance tests because, due to the stratification strategy, grade enrollment is comparable across 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 (90% of the total). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A5: *ITT effect of practice exercises on math achievement at endline, by repeated and non-repeated items*

	(1) Repeated items (proportion-correct) score	(2) Non-repeated items (proportion-correct) score
Treatment	0.005	0.002
	(0.004)	(0.005)
Baseline score	0.089***	0.110***
	(0.004)	(0.005)
N (students)	4001	4001
R-squared	0.527	0.491

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in repeated items across baseline and endline (column 2) and non-repeated items (column 3) domain at endline (after 6 months). Both estimations include randomization-strata fixed effects. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A6: *ITT effect of practice exercises on math achievement at endline, accounting for whether questions included words / academic language*

	A. All students					
	Proportion correct		IRT (separate)		IRT (additional)	
	(1) No words	(2) No acad. lang.	(3) No words	(4) No acad. lang.	(5) No words	(6) No acad. lang.
Treatment	0.000	0.004	0.009	0.013	0.011	0.013
	(0.004)	(0.004)	(0.025)	(0.025)	(0.025)	(0.025)
Baseline score	0.059***	0.090***	0.610***	0.644***	0.597***	0.645***
	(0.004)	(0.004)	(0.025)	(0.025)	(0.025)	(0.025)
N (students)	4001	4001	4001	4001	4001	4001
R-squared	0.434	0.550	0.575	0.588	0.564	0.589
	B. Low-performing students					
	Proportion correct		IRT (separate)		IRT (additional)	
	(1) No words	(2) No acad. lang.	(3) No words	(4) No acad. lang.	(5) No words	(6) No acad. lang.
Treatment	0.017**	0.023***	0.126**	0.132***	0.117**	0.134***
	(0.009)	(0.008)	(0.051)	(0.050)	(0.052)	(0.050)
Baseline score	0.073***	0.099***	0.622***	0.646***	0.610***	0.647***
	(0.007)	(0.007)	(0.043)	(0.042)	(0.043)	(0.041)
N (students)	4001	4001	4001	4001	4001	4001
R-squared	0.436	0.555	0.577	0.590	0.566	0.591

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on math achievement at endline (after six months), accounting for whether questions included words (columns 1, 3, and 5) or academic language (columns 2, 4, and 6). In columns 1 and 2, the outcome is the proportion of students' correct answer for test questions that do not require students to read words (beyond the prompt "solve"), or to read academic terms (e.g., "rectangle"). In columns 3 and 4, the outcome is a (standardized) factor score from a confirmatory, two-dimensional 2PL item response theory model, where items that do not require literacy load on this factor (and items that do require literacy load on another, separate factor that correlates with the first). In columns 5 and 6, the outcome is a (standardized) factor score from a confirmatory, two-dimensional 2PL item response theory model, where all math items load on this factor (and items that require literacy may also load on another, additional factor that correlates with the first). 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. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A7: *Heterogeneous ITT effects on math achievement at endline, by students' sex and enrolled grade*

	Math (IRT-scaled) score	
	(1)	(2)
Treatment	-0.015	0.021
	(0.036)	(0.050)
Female	-0.050	-0.022
	(0.038)	(0.029)
Treatment X Female	0.054	
	(0.050)	
Treatment X Grade 5		-0.011
		(0.068)
Treatment X Grade 6		0.042
		(0.071)
Treatment X Grade 7		-0.067
		(0.073)
Baseline score	0.646***	0.645***
	(0.025)	(0.025)
N (students)	4001	4001
R-squared	0.588	0.588

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after 25 weeks) for female students (column 1) and students enrolled in different grades (column 2). Both estimations include baseline randomization-strata fixed effects. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A8: *ITT effect of practice exercises on usage of CAL platform*

	Number of sessions completed (log)	Total minutes spent on CAL platform (log)	
	(1)	(2)	(3)
A. All students			
Treatment	0.010	0.023	0.015*
	(0.016)	(0.015)	(0.008)
Baseline score	0.068***	0.101***	0.050***
	(0.016)	(0.015)	(0.008)
Number of sessions completed (log)			0.760***
			(0.008)
N (students)	3999	3999	3999
R-squared	0.477	0.504	0.845
B. Low-performing students			
Treatment	0.087**	0.082**	0.016
	(0.036)	(0.033)	(0.018)
Baseline score	0.095**	0.147***	0.075***
	(0.044)	(0.040)	(0.022)
Number of sessions completed (log)			0.756***
			(0.018)
N (students)	979	979	979
R-squared	0.409	0.429	0.824

Notes: (1) 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. (2) The estimations exclude 2 (out of 4,001) students who did not spend any time on the software. (3) Panel A provides results for all students. Panel B provides results for the subsample of students in a grade-level's bottom quartile, as per students' performance on the baseline assessment. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A9: *Robustness of estimation results to standard-error adjustment to account for correlations in student outcomes at the classroom level*

	Main effect		Cont. interact.		By quartile	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.014	0.014	0.014	0.014	0.136***	0.136***
	(0.025)	(0.026)	(0.025)	(0.025)	(0.050)	(0.043)
Baseline score	0.645***	0.645***	0.672***	0.672***	0.647***	0.647***
	(0.025)	(0.027)	(0.028)	(0.031)	(0.042)	(0.056)
Treatment X Baseline			-0.054**	-0.054**		
			(0.025)	(0.026)		
Quartile 2					0.182***	0.182***
					(0.065)	(0.064)
Quartile 3					0.198**	0.198**
					(0.085)	(0.091)
Quartile 4					0.144	0.144
					(0.110)	(0.134)
Treatment X Quartile 2					-0.160**	-0.160**
					(0.071)	(0.065)
Treatment X Quartile 3					-0.131*	-0.131*
					(0.071)	(0.067)
Treatment X Quartile 4					-0.193***	-0.193***
					(0.070)	(0.067)
N (students)	4001	4001	4001	4001	4001	4001
R-squared	0.588	0.588	0.588	0.588	0.590	0.590
Clustered s.e.s	No	Yes	No	Yes	No	Yes

Notes: (1) This table shows the estimations from Tables 2 and 4 (average and heterogeneous ITT effects of practice exercises, respectively) with and without adjusting standard errors to account for correlation in student outcomes at the classroom level. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A10: *Robustness of results to systematic attrition*

	Average effect			Lowest quartile		
	(1) IPW	(2) Lee (lower)	(3) Lee (upper)	(4) IPW	(5) Lee (lower)	(6) Lee (upper)
Treatment	0.015	-0.034	0.060**	0.135***	0.114**	0.134***
	(0.025)	(0.024)	(0.024)	(0.041)	(0.048)	(0.048)
N (students)	4001	3965	3966	4001	3965	3966
R-squared	0.589	0.624	0.609	0.591	0.627	0.610

Notes: (1) This table shows the intent-to-treat (ITT) effect of practice exercises on students' achievement in math at endline (after 6 months). Columns 1-3 show average effects; columns 4-6 show effects among students in the bottom quartile of baseline achievement within their grade level. (2) Columns 1 and 4 show inverse-probability weighted (IPW) estimates; columns 2-3 and 5-6 show Lee (2009) bounds. (3) All estimations include randomization-strata fixed effects and baseline scores (results not shown). Columns 4-6 include, but do not report on, quartile fixed effects and their interactions with the treatment indicator. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.