NBER WORKING PAPER SERIES

DIMINISHING MARGINAL RETURNS TO COMPUTER-ASSISTED LEARNING

Eric Bettinger
Robert W. Fairlie
Anastasia Kapuza
Elena Kardanova
Prashant Loyalka
Andrey Zakharov

Working Paper 26967
http://www.nber.org/papers/w26967

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020, Revised October 2022

We would like to thank Yandex Inc. for data and support for the study. We thank Natalia Lazzati, Jesse Li and Jon Robinson, and seminar participants at UC Berkeley for comments and suggestions. The study was pre-registered at the AEA RCT registry prior to endline data collection. The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'. Approval for the study was obtained from the National Research University Higher School of Economics IRB and Stanford University (IRB #50207). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Diminishing Marginal Returns to Computer-Assisted Learning
Eric Bettinger, Robert W. Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov
NBER Working Paper No. 26967
April 2020, Revised October 2022
JEL No. F63,I25

ABSTRACT

The previous expansion of EdTech as a substitute for traditional learning around the world, the recent full-scale substitution due to COVID-19, and potential future shifts to blended approaches suggest that it is imperative to understand input substitutability between in-person and online learning. We explore input substitutability in education by employing a novel randomized controlled trial that varies dosage of computer-assisted learning (CAL) as a substitute for traditional learning through homework. Moving from zero to a low level of CAL, we find positive substitutability of CAL for traditional learning. Moving from a lower to a higher level of CAL, substitutability changes and is either neutral or even negative. The estimates suggest that a blended approach of CAL and traditional learning is optimal. The findings have direct implications for the rapidly expanding use of educational technology worldwide prior to, during and after the pandemic.

Eric Bettinger
Stanford School of Education
CERAS 522, 520 Galvez Mall
Stanford, CA 94305
and NBER
ebettinger@stanford.edu

Robert W. Fairlie
Department of Economics
Engineering 2 Building
University of California at Santa Cruz
Santa Cruz, CA 95064
and NBER
rfairlie@ucsc.edu

Anastasia Kapuza
National Research University
Higher School of Economics
Moscow
Russia
nas669@yandex.ru

Elena Kardanova
National Research University
Higher School of Economics
Moscow
Russia
e_kardanova@mail.ru

Prashant Loyalka
E413 Encina Hall
Stanford University
Stanford, CA 94305
loyalka@stanford.edu

Andrey Zakharov
National Research University
Higher School of Economics
Moscow
Russia
ab.zakharov@gmail.com

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/4126

1 Introduction

Numerous educational interventions have been used to improve academic achievement and increase human capital among schoolchildren in developing countries. Among these interventions, technology-based interventions have shown promise relative to other popular interventions such as teacher training, smaller classes, and performance incentives (McEwan 2014). It has been argued that EdTech, such as computer-assisted learning (CAL), can offset deficiencies that commonly plague schools, such as low teacher quality, high rates of teacher and student absenteeism, low levels of student motivation, and many students being below grade level, among others (World Bank 2018; Economist 2018; Brookings 2019). These arguments are consistent with the rapid substitution of EdTech for traditional teaching methods and explosion of expenditures on EdTech throughout the world happening even before the pandemic. Furthermore, COVID-19 greatly accelerated these previous trends resulting, at least in the short run, in a whole-scale substitution from traditional learning to EdTech, and a shift to relying on technology especially for home- and after-school work which is likely to persist long after schools return fully to in-class instruction.

The previous findings on the effectiveness of CAL, however, are mixed, ranging from null effects to extremely large positive effects (Bulman and Fairlie 2016; Escueta 2017; Rodriguez-Segura 2021; Abbey et al. 2022). To gain insight into this heterogeneity and add a new dimension of analysis, we design and implement a randomized controlled trial (RCT) involving approximately 6,000 grade 3 students in 343 classes (one per school) from two regions in Russia. The RCT includes three treatment arms: i) CAL for 45 minutes per week, ii) a "double dosage" CAL for 90 minutes per week, and iii) a control that receives no CAL. Estimates of the two treatment effects allow us to explore input substitutability in the use of CAL for the first time in the literature. Importantly, CAL use was directly substituted for traditional learning, avoiding problems associated with identifying separate technology versus increased learning time effects (Ma et al. 2020).

Although extant evidence is from field experiments, heterogeneity in results may stem from variation in the substitutability between CAL and traditional learning. The focus in the previous literature on estimating the average productivity of CAL for a fixed amount of time on CAL provides only limited evidence on characteristics on how this substitutability might change. It does not provide information relevant to important questions regarding input substitutability. In fact,

surprisingly, there is little evidence in the previous literature on the substitutability of any input in the educational production function.[2] Another problem is that evaluating only one level of treatment intensity could be misleading if the level chosen for the experiment is too low or too high relative to other substitutable inputs (i.e. educational production might be suboptimal). Unfortunately, similar to many other inputs in educational production, theory provides only limited guidance on optimal levels of substitution.

This study is the first to discern how the effects of CAL change exogenously with respect to usage levels within the same educational setting.[3] Our study is also one of the only studies that evaluates CAL as a direct substitute for traditional learning instead of being provided as a supplemental after-school program, which likely influences impact estimates. Examining the role of CAL as a direct substitute for traditional learning is also important as countries increasingly mandate limitations on time children spend in after-school programs and on homework.[4] Our use of CAL is also through homework instead of in-class substitution of CAL software. We provide new evidence on the use of CAL for homework. Finally, and perhaps most importantly, direct substitution between the two inputs in the field experiment ensures that any changes in educational production is due to input substitution and not higher inputs. Our study is one of the first to use an experiment to provide evidence on the substitutability of any input in educational production.

We find positive effects of CAL on math test scores at the base dosage level. Doubling the amount of CAL input, we find similar effect sizes relative to the control. We thus find evidence that is consistent with a concave relationship between CAL and educational production. Moving from zero to the base level of CAL, CAL is a positive substitute for traditional learning. But, moving from the base level of CAL to the higher level of CAL, CAL is a similarly productive substitute for traditional learning.

For impacts on language achievement, we find positive effects of CAL at the base level, but stronger evidence consistent with concavity. We find that CAL is a positive substitute when

[2] For example, the one-to-one laptop or home computer programs that have been previously studied do not structure or exogenously determine time use, which is needed to study marginal productivity or input substitutability (e.g. Fairlie and Robinson 2013; Beuermann et al. 2015; Cristia et al. 2017; Hull 2019).

[3] Hypothetically, a meta-analysis of estimates from previous studies could be used to provide evidence on the characteristics of education production, but the CAL programs used in these studies differ by more than usage time (e.g. substitution vs. supplemental program, country, student preparation, grade level, and the presence of additional instructional support).

[4] Policies to reduce time on homework exist, for example, in China (MOE, 2018), France (MNE, 2019), and Russia (SanPiN, 2010).

moving from zero to the base level of CAL, but a negative substitute when moving from the base level of CAL to the higher level. The findings for math and language in CAL do not differ when we shift the focus from mean impacts to impacts throughout the distribution (i.e. quantile treatment effects). We find no evidence of differential treatment effects by gender for either dosage level. For math and language, we do not find clear evidence of differential treatment effects for high-ability students relative to low-ability students.

Our findings contribute to a large literature on the effectiveness of CAL, which provides a wide range of estimates from null effects to extremely large positive effects.[5] Generally, evaluations of supplemental learning CAL programs find large positive effects on academic outcomes (e.g. Lai et al. 2013; 2015; Mo et al. 2014; Bohmer, Burns, and Crowley 2014; Muralidharan et al. 2019; Ito et al. 2019; Araya et al. 2020; Blimpo 2020).[6] For the less common use of CAL as a direct substitute for regular teacher instruction in the classroom or traditional learning after school the evidence often shows null effects (Dynarski et al. 2007, Campuzano et al. 2009; Linden 2008; Barrow et al. 2009; Carillo et al. 2011; Schling and Winters 2018; Taylor 2018; Naik et al. 2020; Ma et al. 2022). Related to these studies of CAL, the less structured provision of computers and laptops for home and/or school use among schoolchildren tends to show null or mixed effects.[7] The findings from our experiment suggest that some of the wide range of estimates on the effectiveness of CAL might be due to chosen dosage levels in addition to study heterogeneity by development level of the country, substitution vs. supplemental program, and features of the software.

The evidence from this analysis helps inform decisions about optimal investment in CAL relative to traditional learning. Identifying optimal levels of investment in CAL is especially important as governments, schools and families are currently investing heavily in EdTech and

[5] See, for example, Banerjee et al. 2007; Linden 2008; Carillo et al. 2011; Angrist and Lavy 2002; Lai et al. 2013, 2015; Mo et al. 2014; Ma et al. 2022; Taylor 2018; Muralidharan et al. 2019; Rouse and Krueger 2004; Dynarski et al. 2007; Barrow et al. 2009; Campuzano et al. 2009; Rockoff 2015; Falck, Mang, and Woessmann 2018; Ito et al. 2019; Araya et al. 2019; Blimpo 2020. Also, see Glewwe et al. (2013), Bulman and Fairlie (2016), Escueta (2017), Rodriguez-Segura 2021, Abbey et al. (2022) for recent reviews of the literature.

[6] Conducting a meta-analysis of the large number of studies conducted in China, Abbey et al. (2022) find that the pooled effect size of the 18 included studies indicates a small, positive effect on student learning (0.13 SD, 95% CI [0.10, 0.17]), and the strongest evidence exists for the effectiveness of CAL that is used as a supplement to existing learning inputs.

[7] See, for example, Malamud and Pop-Eleches 2011; Fairlie and Robinson 2013; Beuermann et al. 2015; Cristia et al. 2017; Malamud et al. 2019; Hull 2019; Fiorini 2010; Fairlie and London 2012; Machin, McNally, and Silva 2007; Schmitt and Wadsworth 2006; Fuchs and Woessmann 2004; de Melo et al. 2014; Yanguas 2020.

likely to increase expenditures in the future. This is especially true for the rapidly growing use of new technologies and their substitution for traditional learning methods in educating schoolchildren in developing countries which was happening prior to COVID and likely has been accelerated because of COVID.

2 Research Design

2.1 Field Experiment

To explore CAL and traditional learning substitutability, we design and implement an RCT involving approximately 6,000 third grade schoolchildren in 343 classes/schools in two provinces of Russia.[8] The RCT includes three treatment arms: an "X" dosage CAL arm where students receive 10 items per subject using the software, which (as communicated to the treatment group) is approximately 20-25 minutes per week of math CAL and 20-25 minutes of (Russian) language CAL; a "2X" dosage CAL arm in which (as communicated to the treatment group) students receive 20 items per subject which is approximately 40-50 minutes of math CAL and 40-50 minutes of language CAL; and a control arm.[9] With this design, we can explore input substitutability in educational production across different levels of CAL.

The field experiment is conducted among primary schools in Russia. Specifically, 343 schools from 2 regions were sampled to participate in the experiment. In each school, one third grade class was sampled, and each class has an average of 18.3 students per class. For each third grade class there is one teacher that teaches both math and language. Altogether, 6,253 students and their 343 teachers were sampled and surveyed.

[8] In some ways, Russia's educational system resembles the educational systems of other OECD countries. The enrollment rate in primary and secondary education is close to 100%. The average class size (21.6 students per teacher in our sample —see Table A1) is also roughly the same as the OECD average for primary schools (21 students per teacher—OECD 2019a). In other ways, however, Russia's educational system is closer to that of other middle-income countries. Its educational expenditures per primary and secondary school student were low at 4,247 US dollars in 2016 (adjusted for purchasing power parity—OECD 2019b). This is less than half the OECD average (9,357 US dollars) and below Chile (5,324 US dollars) and Turkey (4,505 US dollars), but above Mexico (3,062 US dollars—OECD 2019b). Russia's GDP per capita ($10,743 current US dollars in 2017) is further just below Costa Rica (11,677 US dollars), and Maldives (11,151 US dollars) and just above Brazil (9,821 US dollars), China (8,827 US dollars), and Mexico (8,910 US dollars) (World Bank Database 2019). The two regions where the experiment is conducted, Altai Krai (93 schools) and Novosibirsk (250 schools), have GDP per capita below the national average (OECD 2019b).

[9] Unfortunately, the company was unable to provide complete data on CAL usage across the Dosage 1X and Dosage 2X groups (which was a goal for data collection stated in our pre-analysis plan). Interviews with teachers revealed that they generally complied with instructions on use, which is consistent with bi-weekly follow-ups by the provider on usage of the software.

In the second half of October 2018 (near the start of the school year), we conducted a baseline survey of the sampled students, their teachers and principals. After the baseline survey, we randomized classes to treatment conditions. Students participated in the treatment from December 2018 until mid-May 2019. In mid May 2019, the end of the Russian school year, we administered a follow-up survey with students, teachers, and principals.

2.2 CAL

The provider of the CAL software is the largest technology company in Russia (hereafter "the provider"). The provider's platform has more than 10,000 items across various math and language sub-content areas for grades 2 to 4. The items and associated content areas align closely with national educational standards and curricula for primary schools, and thus the problems are similar to those in traditional assignments. As such, the platform was intended to be used throughout the country. After our evaluation, it was widely adopted by schools in many regions.

The CAL software is of high quality and similar to that used in previous studies. It has a graphics-based and attractive user-interface and dynamic, engaging tasks. It allows multiple tries per question and provides scaffolded feedback after each student response. The software also allows teachers to track and compare the performance of individual students both overall and at a granular level in subject-specific content and sub-content areas. Appendix A presents example screenshots of these different aspects of the CAL software.

Students in the treatment group use CAL at home as a partial or full replacement for traditional pencil and paper homework. Traditionally, teachers give students a certain number of homework exercises in class, ask students to complete the assigned exercises at home (using pencil and paper), and then finally turn in the completed exercises in class. For the treated students, some or all of the traditional pencil and paper exercises are replaced by time on CAL. Homework, whether traditional or replaced by CAL, reviews concepts and allows students to practice and solidify their knowledge of what was learned in class.

2.3 Baseline Survey

We administered three baseline surveys to students and teachers. The student survey collected basic background information such as student gender and time spent on math and language homework. As part of the baseline survey, we administered proctored exams in four areas: math,

language, reading, vocabulary (math and language achievement were our pre-determined main academic outcomes).[10] As noted in Appendix B, the exams have good psychometric properties. The teacher survey further collected information on the degree to which teachers use information and computing technology (ICT) at home and their self-efficacy with ICT.

2.4 Randomized Design and Statistical Power

To maximize statistical power, we created the sample strata or blocks by placing the six classes with the closest mean grade three math scores in a region in a strata.[11] Adjusting for strata, the resulting intraclass correlation coefficients were extremely low for our two main outcomes: 0.000 in math achievement and 0.053 in language achievement. Classes were then randomly allocated within strata (conducting randomization once) to one of three different treatment conditions (T1 = CAL Dosage 1X, T2 = CAL Dosage 2X, or C = Control or No CAL):

A. CAL Dosage 1X (T1)	115 classes (in 115 schools)
B. CAL Dosage 2X (T2)	113 classes (in 113 schools)
C. Control (C)	115 classes (in 115 schools)

The large number of schools per treatment arm, extremely low ICCs, and rich set of baseline controls provide substantial statistical power with which to measure effects.[12] Even without controlling for baseline test scores, minimum detectable effect sizes (MDESs) are approximately 0.09 SDs (for math) and 0.12 SDs (for language) for pairwise treatment comparisons.

[10] Details of the baseline data collection (and proposed analyses) were described in a pre-analysis plan written and filed with the American Economic Association registry before endline data were available for analysis (https://www.socialscienceregistry.org/trials). Due to minor technical difficulties in the baseline survey (before randomization), not all 6,253 students took all four tests. Rather, 6,052 students in the baseline took math and vocabulary tests, while 5,838 students took language and reading tests. We deal with missing values for these and other baseline controls by including missing value dummies (as detailed in the pre-analysis plan).

[11] Because the number of schools in each region was not divisible by 6, we placed 9 schools (with the closest mean grade 3 math scores) in the first region in one stratum and 10 schools (with the closest mean grade 3 math scores) in the second region in one stratum.

[12] Based on a previous longitudinal study in primary schools in Russia using the same test instruments, the estimated R-squared between the baseline and follow-up scores is approximately 0.50. Other parameters for the power calculation include: 18 students per class/school, an alpha of 0.05 and power = 0.8.

2.5 Balance Checks

Appendix Table A1 presents summary statistics for the baseline variables as well as tests for balance on baseline observables across the treatment arms. The exam scores are standardized as z-scores and thus have a mean of 0 and standard deviation of 1. The percentage of students that are female is 52% and the average class size is 21.6 students. The table also shows the results from a total of 24 tests comparing average variable values among the treatment and control arms. These tests were conducted by regressing each baseline variable on a treatment group indicator and controlling for strata. For tests of student-level variables, standard errors are clustered at the school level.

Out of the 24 tests, only one was statistically significant (different from zero) at the 10% level and none were significant at the 5% or 1% levels. The results from Table A1 indicate that balance was achieved across the three arms, especially as a small number of significant differences are to be expected (by random chance). A joint test of all baseline covariates simultaneously shows no significant difference between T1 and C (p-value: 0.265), T2 and C (p-value: 0. 178) or T1 and T2 (p-value = 0.160). Key baseline covariates (baseline math and language test scores, not to mention reading and vocabulary scores) were not statistically different between any of the three treatment arms (even at the 10% level) with only one exception. There is a slight imbalance in the baseline math score between Dosage 2X vs control, but the difference is only marginally significant, the difference is negative working against finding a dosage 2X vs control effect, and we control for it in the treatment regressions.

2.6 Program (Treatment) Administration

In both the CAL Dosage 1X and CAL Dosage 2X treatment arms, the provider asked teachers to assign CAL items to their classes through their registered accounts.[13] Teachers were given instructions to use assigned CAL items during homework but were also allowed to use them in class.[14] The vast majority of teachers reported using CAL for homework (more than 95%).

One reason that increasing the dosage of CAL could result in increased effectiveness is that it might have increased *total* time on homework. Conversely, if there was crowd-out (i.e. the

[13] The dosages were chosen based on numerous pilot interviews that the provider conducted with teachers outside of the study sample and prior to the experiment. In the experimental intervention, the provider introduced the online educational platform and dosages through separate training webinars with the Dosage X and Dosage 2X teachers.
[14] Interviews with teachers revealed that class use was minimal relative to use for homework.

substitution between CAL and traditional learning was less than one) then we could find a decrease or no increase in effectiveness. To explore this question, we examine total hours spent on homework by students by treatment condition. Table 1 reports estimates of total homework hours on math and language from regressions with only baseline score controls and with baseline score plus additional controls. Although reported hours might be somewhat underreported the comparisons are informative. We find precise zero estimates, indicating that, compared to the control condition (mean=43 for math and mean=44 for language), neither CAL treatment condition (Dosage 1X or Dosage 2X) resulted in greater or lower total time on homework in either subject (as reported by students).[15] The frequency of homework assignments also did not differ significantly among the treatment control groups. Qualitative interviews further indicate that teachers almost always substituted (instead of supplemented) traditional learning activities with CAL.[16] Teachers in the treatment conditions also did not change the amount of time they prepared for their math and language lessons relative to the control group (Table 2). We thus treat CAL and traditional learning as being substituted one-to-one when interpreting our results.[17]

The dosages of CAL are in line with those used in recent studies. For example, Lai et al. (2013; 2015) and Mo et al. (2014) find large positive effects of supplemental CAL programs for Chinese schoolchildren (0.12 to 0.18σ in math) from 40 minutes of instruction, 2 times a week. Some studies use larger dosages. Bohmer, Burns, and Crowley (2014) find large positive effects from an after-school program providing CAL and student coaches in South Africa (0.25σ in math) from 90 minutes twice a week, but part of the program includes student coaches. Banerjee, Cole, Duflo and Linden (2007) find that 120 minutes per week of CAL improves grade 4 math test scores by 0.35 SDs after one year. Muralidharan et al. (2019) find large positive effects of after-school Mindspark Center programs in India which include software use and instructional support (0.59σ in math and 0.36σ in Hindi) from 90 minutes per session, six sessions a week. However, requiring schoolchildren to use CAL in addition to pre-existing homework at these much higher levels is just not possible in most countries. As noted above, many countries mandate limitations on time

[15] Distributions of total homework time align almost perfectly for the control, Dosage 1X and Dosage 2X groups.
[16] When asked directly about whether they assigned more homework to their class as a result of the intervention, the vast majority of interviewed teachers said no. It was also clear from pilot interviews that teachers were highly sensitive to assigning additional homework to students because the law sets limits on the total amount of homework time that can be assigned to students (1.5 hours per day in all subjects—SanPiN 2010).
[17] We unfortunately do not have data about teaching styles and are thus unable to examine whether the interventions changed teaching styles.

children spend in after-school programs and on homework (e.g. China (MOE, 2018); France (MNE, 2019); and Russia SanPiN, 2010); in the United States many school districts have already or are considering implementing homework restrictions (Tawnell, 2018).

2.7 Endline Survey and Primary Outcomes

We conducted the follow-up survey with students and teachers in mid May 2019 at the end of the school year. As in the baseline, we administered a 2-hour proctored exam that covers math, language, reading, and vocabulary to students. Proctors were independent from the schools. They were recruited from regional universities and educational policy organizations, such as regional centers of educational assessment. School workers were not allowed to be proctors. We also asked students about their homework time on different subjects, and we asked teachers about their preparation time for teaching different subjects.

The primary outcome variables for the trial are student math and language achievement at the end of the school year (as measured by the exam). In the analyses, we convert the math and language endline exam scores (percent correct) into z-scores (subtracting each students' endline subject-specific score by the average endline subject-specific score of the control sample and dividing the standard deviation of the endline subject-specific score of the control sample). Other outcome variables include the degree to which students are interested in studying math and language subjects (using a standard subjective scale, converted into z-scores), student reports of time spent on subject-specific homework (average minutes per week), and teacher reports of time spent preparing for teaching different subjects (average minutes per week).[18]

3 Empirical Methods and Hypothesis Tests

[18] Out of the baseline sample of 6,052 students that took the math test in the baseline, 5,552 students (92%) took the math test in the endline; an additional 165 students took math in the endline but not in the baseline. Out of the baseline sample of 5,838 students that took the language test in the baseline, 5,205 students (89%) took the language test in the endline; an additional 360 students took language in the endline but not in the baseline. The missingness rates for the math and language analytical samples were 8% and 11% respectively. Balance in baseline covariates across pairwise treatment comparisons was maintained among the non-missing students. Out of 24 tests, only two were statistically significant (different from zero) at the 10% level and none were significant at the 5% or 1% levels (Appendix Table A2), as would be expected by chance.

Our general approach for estimating treatment effects is to regress math and language outcomes on indicator variables for treatment assignment, baseline controls and strata fixed effects using the following model:

$$Y_{ij} = \alpha + \gamma_1 D_{1j} + \gamma_2 D_{2j} + X_{ij}\beta + \tau_s + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the outcome of interest measured at endline for student *i* in school *j*; D_{1j} and D_{2j} are dummy variables indicating the treatment assignments of Dosage 1X and Dosage 2X; X_{ij} is a vector of baseline control variables, and τ_s is a set of strata fixed effects.[19] In all specifications, X_{ij} includes the baseline value of the dependent variable (when available). We also estimate treatment effects using an expanded set of baseline controls. For student-level outcomes, this expanded set of baseline controls includes all baseline test scores (math, language, reading, and vocabulary), student gender, an indicator for whether the teacher uses ICT at home, teacher ICT self-efficacy, and class size.[20] Standard errors are clustered at the school level.

The key parameters of interest in Equation (1) are γ_1 and γ_2. These estimates shed light on whether the production function is concave in CAL. For example, the finding of a positive estimate of γ_1 and an estimate of γ_2 that is less than $2\gamma_1$ is consistent with a concave relationship. Estimates of γ_1 and γ_2 also allow one to determine if substitution between the CAL and traditional learning inputs increases academic achievement. We can specifically examine whether substitutability diminishes with higher levels of CAL. Having three treatment arms of different dosage (including the control arm where dosage is zero) in the RCT allows us to explore these questions.

We chose the Dosage 1X and Dosage 2X levels of CAL use because, as noted above, they fall within the range of what teachers believe are reasonable amounts, are within policy regulations, and line up well with levels implemented in the previous literature. Another important point of the experimental design is that we are increasing CAL by substituting away from traditional learning which is different than adding a supplemental CAL program. This allows us to isolate changes in educational production resulting from input substitution instead of productivity changes due to

[19] As pre-specified in our pre-analysis plan, we focus on math and language outcomes. Our primary outcomes are math and language achievement as measured by standardized test scores. Course grades for students were not available from all schools.

[20] We address missing values for the baseline controls by creating a missing value dummy variable and including it in the regression.

changing input levels. This is an important distinction because schools and students face restrictions on in-school and after-school time commitments.

4 Results

4.1 Math Scores

Table 3 reports estimates of math test scores on treatment arms. Both specifications with only baseline score controls and with baseline score plus additional controls are reported. For Dosage 1X we find positive and statistically significant effects on math test scores (0.10 to 0.11σ). Using CAL increased test scores and the increase at the base level of time resulted in effect sizes that are roughly comparable to estimates reported in previous studies at similar dosage levels. For example, Lai et al. (2013; 2015) and Mo et al. (2014) find 0.12 to 0.18σ effects in math from CAL programs for Chinese schoolchildren from 80 minutes per week.

After doubling the dosage level, we also find positive and statistically significant treatment effects on math test scores. More importantly, however, we find point estimates that are similar to the first dosage level. Increasing the dosage level thus resulted in no additional increase in effects on math test scores. To our knowledge, these estimates are the first showing no additional effect of a higher dosage of CAL beyond the base level.

One question of interest is whether production in CAL is concave.[21] In this case, the positive substitutability of CAL for traditional learning diminishes as CAL is expanded beyond roughly equal levels. As CAL use is expanded, one possibility is that each additional unit becomes less productive because students become less interested or engaged in the graphics- and video-based learning with more use. Another possibility is that higher levels of CAL use increase the likelihood that students become distracted with other software, apps and entertainment on the computer. On the other hand, production in CAL might not be concave. An example of this case might be that CAL and traditional learning are perfect substitutes for each other across all levels.[22]

Having three treatment arms of different dosage (including the control arm where dosage is zero) in the RCT allows us to explore this question empirically for the first time. We first examine whether the estimates are consistent with concavity by comparing the impact of the 2X

[21] A standard Cobb-Douglas production function with equal factor returns, for example, implies concavity because of the curvature in isoquants.
[22] A linear production function in which both inputs have similar returns, for example, implies non-concavity.

dosage to 2 times the impact of the 1X dosage (where both impacts are relative to the control). Table 3 reports the results of the test. We find some limited evidence that is consistent with concavity in educational production in CAL.

Turning to the implications for the substitutability between CAL and traditional learning, the estimates of the two treatment effects suggest different substitutability depending on the base level of CAL. We find that moving from zero to the lower level of CAL, the substitutability of CAL for traditional learning is greater than one (i.e. traditional learning can be reduced by more than one unit when CAL is increased by one unit), but moving from the lower level of CAL to the higher level of CAL the substitutability is equal to one (i.e. CAL and traditional learning are perfectly substitutable across this range). These findings also provide some suggestive evidence on the general forms of the educational production function as discussed in Bettinger et al. (2021).

The test of two different levels of CAL is also useful beyond testing for concavity or examining input substitutability. For example, testing for the positive effect of each CAL dosage is of immediate interest to the CAL provider (the largest technology company in Russia) as well as to local and national policymakers in Russia (since, to the best of our knowledge, this is the first randomized evaluation of EdTech in Russia). Evaluating only one level of treatment intensity could be misleading for identifying whether CAL is effective if the level chosen for the experiment is too low or too high. We find positive and statistically significant effects for both treatment levels suggesting that different choices of levels of CAL can improve math test scores.

4.2 Language Scores

We also examine treatment effects on language test scores. The previous literature focuses more on math test scores than on language test scores. Languages differ in each country making it difficult to choose base levels and compare estimates across studies. Additionally, we might expect that educational production in CAL differs between math and language. Although math learning is mostly through school and homework, language learning is broader because reading for pleasure and family interactions also play key roles in learning.

Table 3 reports estimates for language test scores. Both specifications with only baseline score controls and with baseline score plus additional controls are reported. For Dosage 1X we find some evidence of positive and statistically significant effects (at the 0.10 level) on language test scores. After doubling the dosage level, the treatment effect estimates become close to zero.

Table 3 also reports the estimates that provide suggestive evidence on the concavity test. For impacts on language achievement we find positive effects of CAL at the base level, but stronger evidence that is consistent with concavity in the production function. We find a positive substitutability of CAL for traditional learning moving from zero to the lower level of CAL, but a negative substitutability moving from the lower level of CAL to the higher level. If the experiment had only estimated the treatment effect at the higher dosage level in CAL, the positive effects at the lower level, curvature, and changing substitutability would have been missed.

The findings clearly indicate that there is an optimal amount of CAL use for language that represents a relatively balanced approach instead of one with very high levels of usage (or no usage). Additionally, if the experiment only provided the higher dosage of CAL then it would have concluded with a clear null effect on language test scores. This represents a more general concern in tests of the effectiveness of CAL that rely on only one input level.

4.3 Interest in Studying Math and Language

A common argument for how CAL, or EdTech more generally, works is that it increases interest to engage with subject material. If students enjoy learning math, for example, through CAL that enjoyment could spill over to learning math more generally. Thus, one reason that substituting CAL for traditional learning at the base level might increase math achievement is because CAL engages kids and encourages them to study math through its graphics and gamified nature. Additionally, any curvature in isoquants could be partly due to diminishing engagement in math as CAL is increased relative to traditional learning. Diminishing engagement could be due, for example, either to limited attention spans (that benefit from a mix of traditional and computer-based homework) or greater fatigue (because of the more intense, interactive nature of the CAL exercises).

Table 4 reports estimates of Equation (1) for whether students are interested in studying math and language. The questions underlying the measure do not refer to CAL and are more generally focused on interest in math or language. At the base dosage level the math interest of the treatment group is 0.09σ higher than the control group. Moving to the higher dosage level in CAL, the point estimates become smaller and lose statistical significance from the control, but are not statistically different from the Dosage 1X estimates. Although these results are only suggestive, they are consistent with the lower use of CAL increasing interest more generally in math and thus

resulting in higher math test scores. But, when using CAL more extensively and traditional learning consequently less, students might have become less interested and motivated in math and thus experienced no resulting increase in math test scores. These patterns are consistent with the concave educational production function in CAL and related curvature in isoquants.

The patterns are also strong for interest in studying language. We find large positive estimates from the lower dosage of CAL. Interest to study language increases by 0.08-0.09σ relative to the control. Doubling the dosage of CAL results in a small negative to no change in interest relative to the control. These estimates are consistent with the findings for language test scores and are consistent with more concavity in CAL and curvature in isoquants when we focus on language relative to math.

4.4 Distributional Effects

The results from the treatment regressions provide evidence of CAL effects at the mean. Turning the focus to other parts of the distribution, we estimate quantile treatment effects regressions to test for differential treatment effects across the post-treatment outcome distribution. Appendix C: Figures 1 and 2 display estimates and 95 percent confidence intervals for each percentile for the Dosage 1X and Dosage 2X effects for math and language test scores, respectively. For math test scores we find some evidence that treatment effects are larger in the middle and top of the distribution than the bottom of the distribution. For most of the distribution we find positive and similar-sized estimates of Dosage 1X and Dosage 2X effects (except possibly at the very top of the distribution where there is more noise).

For language scores, the patterns are consistent with the findings for mean treatment effects. Dosage 1X has positive effects throughout the distribution, whereas Dosage 2X has no effects. Although the quantile treatment estimates are not as precisely measured they do not change the conclusion from the mean impacts reported in Table 3. Mean impact estimates do not appear to be concealing differential effects at different parts of the distribution.

4.5 Heterogeneous Effects

We next examine heterogeneous effects by two important subgroups. We focus on differences based on gender and baseline ability (above and below the median). Treatment effects might differ by gender because boys and girls use computers differently with much higher levels of video game

use among boys (Kaiser Family Foundation 2010; U.S. Department of Education 2011; Fairlie 2017; Algan and Fortin 2018). Exploring heterogeneity by baseline ability might be important because, for example, lower ability students might have more room to make gains in test scores than high ability students from using CAL, or lower ability students might benefit more from engaging video-based and gamified instruction. Differences might not reveal when focusing on one treatment level (i.e. average productivity at that point) and instead might manifest in degrees of concavity.

Appendix Tables A3 and A4 report estimates of interactions by gender on achievement and interest in subject, respectively. As expected, we find evidence that girls have higher language test scores than boys, but similar levels of test scores in math (see OECD 2019a, for example). However, even with the difference in language scores, we do not find evidence of differential treatment effects by gender at either Dosage 1X or Dosage 2X for either math or language. The estimates for interest in math and language also show higher interest in language among girls than boys, but no differences in math interest or dosage effects by gender.

We next examine differences by baseline ability level. Appendix Tables A5 and A6 report estimates of interactions between the Dosage 1X and Dosage 2X treatments, and above median baseline ability for test scores and interest in the subject, respectively. For math, the main treatment effects are positive and significant for both dosage levels. The point estimates for the difference in Dosage 1X vs control treatments effects between the bottom and top half of students are not statistically significant. The point estimates for the difference in Dosage 2X vs control effects between the bottom and top half of students show some evidence of marginal significance. For language, we find little statistically significant evidence of positive or negative effects for main effects. We find only limited evidence of a positive differential Dosage 2 vs control effect between students in the bottom and top half of the baseline language ability distribution). For liking subjects the estimates are noisier but generally line up with the test score results. Overall, we do not find clear evidence of differential treatment effects for high-ability students relative to low-ability students in either test score.

5 Conclusion

Billions of dollars are spent on computer-based learning in schools in developing countries each year and substantially more has been spent from the accelerated shifts to technology to facilitate

remote learning resulting from the pandemic, but what are the effects of this massive shift towards EdTech? Unfortunately, there is limited theoretical guidance on what optimal levels of CAL should be, and the newness of EdTech in developing countries does not provide a long enough track record to determine what works, what does not work, and what are the impacts of the continued substitution of CAL for traditional learning. The empirical evidence, even from RCTs, is decidedly mixed and focuses exclusively on one dosage level in CAL. To remedy this deficiency in the literature, we study for the first time the effectiveness of CAL on the educational outcomes of school children at different levels of treatment intensity, which sheds light on the substitutability of CAL for traditional learning. Our field experiment involving approximately six thousand Russian schoolchildren and three treatment arms varying dosage levels in CAL generates exogenous variation in CAL use. CAL is substituted directly for traditional learning through homework in the experiment. The experiment provides novel evidence on the substitutability of inputs not only for CAL but for any input in the educational production function.

Estimates from the field experiment indicate that CAL increases math test scores at both the base and higher dosage levels. As traditional learning is substituted for CAL from the base level to the higher level, however, we find similar effect sizes. Taken together, this suggests that the substitutability of CAL for traditional learning is positive when moving from zero to the base level of CAL, but is neutral when moving from the base level of CAL to the higher level of CAL. These estimates are consistent with CAL having a positive return in educational production. Turning to language achievement, which has been studied less in the previous literature, we find stronger evidence of diminishing substitutability of CAL for traditional learning. The experimental estimates for the returns to CAL for language depend on the level chosen. Importantly, if the experiment had only estimated the treatment effect at the higher dosage level in CAL, the positive effects at the lower level would have been missed. Better knowledge about substitutability is important especially as the widespread substitution of EdTech that happened around the world due to COVID is not likely going away entirely as we move towards more blended approaches in the future.

A novel and important finding is that educational production does not appear to fit a situation in which teachers and students can simply substitute between CAL and traditional learning at any level with the same result. For both math and language achievement we find evidence of diminishing MRTS of CAL for traditional learning. The marginal costs of shifting

from a lower level to a higher level of CAL are very low because students already have computers and the software is online and can be replicated for essentially no cost. Although there are fixed costs of developing the software and keeping it up-to-date, the provider made it free of charge to all schools and teachers in the country. In any case, we do not expect that costs will shift the optimal levels much beyond what we find without detailed measures of costs. The primary constraint in this setting is total homework time mandated by the government.

Why do we find evidence of diminishing substitutability between CAL and traditional learning? One possibility that is at least consistent with our experimental findings is based on changes in interest and engagement in the subject matter. We find that for both math and language, the base level of CAL resulted in the highest levels of interest. When the dosage level of CAL was doubled students reported lower levels of interest. The finding of diminishing substitutability might be due to these effects on interest and engagement in subject material. Another possibility is that at base level dosages of CAL students gain from being more engaged in learning the material through the technology, but at higher dosages they lose out on the positive effects of traditional learning. In the end, a blended approach might be the optimal solution for schools and students. The blended approach might keep students engaged, but at the same time expose students to more beneficial methods of learning or just keep students switching around. The full-scale switch from in-person instruction to online instruction due to COVID is a good example of potential negative impacts on engagement.

More research is needed on these important underlying questions regarding how students learn using technology and more broadly on the substitutability of other educational inputs. Findings from future research along these lines will build on the novel findings presented here on substitutability of CAL for traditional learning and help further identify optimal levels of investment in CAL, which is imperative as governments, schools and families around the world were increasing investments in EdTech and substituting EdTech for traditional learning methods even prior to the greater movement towards EdTech in response to COVID.[23] And, the shift to relying on technology especially for home- and after-school work is not likely to return to pre-

[23] The COVID pandemic, however, does not provide a good natural experiment for examining the effects of substituting towards EdTech because too many other factors changed at the same time (Bacher-Hicks and Goodman 2021). For examples of research examining the broad impacts of COVID on educational outcomes see, for example, Bird, Castleman, and Lohner (2022), Altindag, Filiz, and Tekin (2021), Kofoed, et al. (2021), and Bulman and Fairlie (2022).

pandemic levels, but instead increase to higher levels even after schools return to in-class instruction.

References

Abbey, Cody, Yue Ma, Muizz Akhtar, Dorien Emmers, Robert Fairlie, Ning Fu, Hannah Faith Johnstone, Prashant Loyalka, Scott Rozelle, Hao Xue, and Xinwu Zhang. 2022. “EdTech Innovations and K-12 Student Learning Outcomes in China: A Systematic Review and Meta-analysis” Stanford University REAP Working Paper

Algan, Yann, and Nicole M. Fortin. 2018. "Computer Gaming and the Gender Math Gap: Cross-Country Evidence among Teenagers." In *Transitions through the Labor Market: Work, Occupation, Earnings and Retirement*, pp. 183-228. Emerald Publishing Limited.

Altindag, Duha Tore, Elif S. Filiz, and Erdal Tekin. 2021. *Is Online Education Working?* National Bureau of Economic Research Working Paper No. w29113.

Angrist, Joshua, and Victor Lavy. 2002. "New evidence on classroom computers and pupil learning." *The Economic Journal* 112.482: 735-765.

Auriol, E. and Warlters, M., 2012. The Marginal Cost of Public Funds and Tax Reform in Africa. *Journal of Development Economics*, 97(1), pp.58-72.

Bacher-Hicks, Andrew, and Joshua Goodman. 2021. “The Covid-19 Pandemic Is a Lousy Natural Experiment for Studying the Effects of Online Learning” *Education Next*, Fall, 38-42.

Bai, Y., Tang, T., Wang, B., Mo, D., Zhang, L., Rozelle, S., Auden E., and Mandell, B. 2018. “Impact of Online Computer Assisted Learning on Education: Evidence from a Randomized Controlled Trial in China,” *REAP Working Paper.*

Banerjee, A., Cole, S., Duflo, E. and Linden, L. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics* 122(3): 1235-1264.

Barrow, L., Markman, L. and Rouse, C.E. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction," *American Economic Journal: Economic Policy* 1(1): 52-74.

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

Bettinger, Eric, Robert Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Kumar Loyalka, and Andrey Zakharov. 2021. "Does EdTech substitute for traditional learning? Experimental estimates of the educational production function." NBER Working Paper w26967.

Bird, Kelli, Benjamin L. Castleman, and Gabrielle Lohner. 2020. "Negative impacts from the shift to online learning during the COVID-19 crisis: evidence from a statewide community college system." Annenberg Institute for School Reform at Brown University. EdWorkingPaper 20-299.

Blimpo, M. P., Gajigo, O., Owusu, S., Tomita, R., & Xu, Y. (2020). “Technology in the classrrom and learning in secondary schools.” Unpublished manuscript.

Böhmer, B., Burns, J., and Crowley, L. 2014. "Testing Numeric: Evidence from a Randomized Controlled Trial of a Computer Based Mathematics Intervention in Cape Town High Schools.” Unpublished manuscript.

Brookings 2016. “Classroom technologies narrow education gap in developing countries,” Steven Livingston, August 23, 2016, https://www.brookings.edu/blog/techtank/2016/08/23/classroom-technologies-narrow-education-gap-in-developing-countries/

Bulman, G., and Fairlie, R.W. 2016. "Technology and Education: Computers, Software, and the Internet," *Handbook of the Economics of Education*, Vol. 5, eds Eric Hanushek, Steve Machin, and Ludger Woessmann, North-Holland, Chapter 6: 239-280.

Bulman, George, and Robert W. Fairlie. 2022. The Impact of COVID-19 on Community College Enrollment and Student Success: Evidence from California Administrative Data. National Bureau of Economic Research Working Paper No. w28715.

Burguillo, J. C. 2010. Using Game Theory and Competition-based Learning to Stimulate Student Motivation and Performance. *Computers & Education*, 55(2): 566–575.

Campuzano, L., M. Dynarski, R. Agodini, K. Rall, and A. Pendleton. 2009. Effectiveness of reading and mathematics software products: Findings from two student cohorts." Unpublished manuscript. Washington, DC: Mathematica Policy Research.

Carrillo, P., Onofa, M., and Ponce, J. 2010. "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." *Inter-American Development Bank Working Paper*.

Comi, Simona Lorena, Gianluca Argentin, Marco Gui, Federica Origo, and Laura Pagani. 2017. "Is it the way they use it? Teachers, ICT and student achievement." *Economics of Education Review* 56: 24-39.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2017. "Technology and child development: Evidence from the one laptop per child program." *American Economic Journal: Applied Economics* 9, no. 3: 295-320.

de Melo, G., Machado, A., & Miranda, A. (2014). “The Impact of a One Laptop per Child Program on Learning: Evidence from Uruguay.” IZA, Discussion Paper No. 8489.

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, B. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, and W. Sussex. 2007. Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort." Unpublished manuscript. Washington, DC: Mathematica Policy Research.

Ebner, M. & Holzinger, A. 2007. Successful Implementation of User-centered Game Based Learning in Higher Education: an Example from Civil Engineering. *Computers & Education*, *49*, 3, 873–890.

Economist. 2018. In poor countries technology can make big improvements to education, Economist, November 17, 2018. https://www.economist.com/international/2018/11/17/in-poor-countries-technology-can-make-big-improvements-to-education.

Escueta, M., Quan, V., Nickow, A.J. and Oreopoulos, P. 2017. "Education Technology: An Evidence-Based Review." *NBER Working Paper w23744*.

Fairlie, Robert W. 2016. "Do Boys and Girls Use Computers Differently, and Does it Contribute to Why Boys Do Worse in School than Girls?." *The BE Journal of Economic Analysis & Policy* 16.1: 59-96.

Fairlie, Robert W., and Rebecca A. London. 2012. "The effects of home computers on educational outcomes: Evidence from a field experiment with community college students." *The Economic Journal* 122.561: 727-753.

Fairlie, R.W., and 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.

Falck, O., Mang, C., and Woessmann, L. 2018. "Virtually No Effect? Different Uses of Classroom Computers and their Effect on Student Achievement." *Oxford Bulletin of Economics and Statistics* 80(1): 1-38.

Fiorini, Mario. 2010. "The effect of home computer use on children's cognitive and non-cognitive skills." *Economics of Education Review* 29.1: 55-72.

Fogelholm, M. et al. 2015. "Correlates of Total Sedentary Time and Screen Time in 9–11 Year-old Children around the World: the International Study of Childhood Obesity, Lifestyle and the Environment." *PloS One* 10(6): e0129622.

Fuchs, Thomas, and Ludger Woessmann. 2004. "Computers and Student Learning: Bivariate and Multivariate Evidence on the Availability and Use of Computers at Home and at School." CESIFO Working Paper No. 1321.

Glewwe, Paul W., Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina. 2013. "School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010," in *Education Policy in Developing Countries* (ed. Paul Glewwe): University of Chicago Press: Chicago.

Hobbs, T.D. "Down With Homework, Say U.S. School Districts," *The Wall Street Journal*, December 12, 2018, available at https://www.wsj.com/articles/no-homework-its-the-new-thing-in-u-s-schools-11544610600.

Hull, Marie, and Katherine Duch. 2019. "One-to-One Technology and Student Outcomes: Evidence From Mooresville's Digital Conversion Initiative." *Educational Evaluation and Policy Analysis* 41.1: 79-97.

Ito, Hirotake, Keiko Kasai, and Makiko Nakamuro. 2019. "Does computer-aided instruction improve children's cognitive and non-cognitive skills?: Evidence from Cambodia." Research Institute of Economy, Trade and Industry (RIETI) working paper.

Kaiser Family Foundation. 2010. Generation M2: Media in the Lives of 8- to 18-Year Olds. Kaiser Family Foundation Study.

Kofoed, Michael, Lucas Gebhart, Dallas Gilmore, and Ryan Moschitto. 2021. "Zooming to Class?: Experimental Evidence on College Students&Apos; Online Learning During Covid-19." IZA Discussion Paper 14356.

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.

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., ... & 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.

Levin, H.M. and Belfield, C. 2015. "Guiding the Development and Use of Cost-effectiveness Analysis in Education." *Journal of Research on Educational Effectiveness* 8(3): 400-418.

Li, Y.Y., Li, G.R., Liu, C.F., Loyalka, P., Rozelle, S. 2019. Learning Trajectories among Middle School Students in Developing Contexts: Evidence from Rural China." *REAP Working Paper.*

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

Ma, Y., Fairlie, R., Loyalka, P. Rozelle, S., Bai, Y. 2022. "Identifying the "Tech" in EdTech: Experimental Evidence on Computer Assisted Learning in China" Working Paper.

Machin, Stephen, Sandra McNally, and Olmo Silva. 2007. "New Technology in Schools: Is There a Payoff?" Economic Journal, 117(522): 1145-1167, July.

Malamud, Ofer, Santiago Cueto, Julian Cristia, and Diether W. Beuermann. "Do children benefit from internet access? Experimental evidence from Peru." *Journal of Development Economics* 138 (2019): 41-56.

Malamud, Ofer, and Cristian Pop-Eleches. "Home computer use and the development of human capital." *The Quarterly journal of economics* 126.2 (2011): 987-1027.

McClelland, G.H. 1997. "Optimal Design in Psychological Research." *Psychological Methods* 2(1): 3-19.

McEwan, P. J. 2015. Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments. *Review of Educational Research*, 85(3): 353-394.

Ministry of Education, 2011. 2010 National Statistical Report on Education Development. *China Geology Education*, 2011(3): 93-96.

Ministry of Education. 2018. Notice of the Ministry of Education and Nine Other Departments on the Issuance of Burden Reduction Measures for Primary and Secondary School Students. http://www.moe.gov.cn/srcsite/A06/s3321/201812/t20181229_365360.html

MNE. (2019). Encouraging student success: Homework Done. www.education.gouv.fr. Retrieved from http://www.education.gouv.fr/cid131710/encouraging-student-success-homework-done.html.

Mo, D., Huang, W., Shi, Y., Zhang, L., Boswell, M., & Rozelle, S. 2015. Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China. *China Economic Review*, 36: 131-145.

Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M. and Rozelle, S., 2013. Can One-to-one Computing Narrow the Digital Divide and the Educational Gap in China? The Case of Beijing Migrant Schools. *World Development*, 46: 14-29.

Mo, D., Zhang, L.X., Luo, R.F., Qu, Q.H., Huang, W.M., Wang, J.F., Qiao, Y.J., Boswell, M., and 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(3): 300-323.

Muralidharan, K., Singh, A., and Ganimian, A.J. 2019. Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review* 109(4): 1426-60.

Naik, Gopal, Chetan Chitre, Manaswini Bhalla, and Jothsna Rajan. 2020. "Impact of use of technology on student learning outcomes: Evidence from a large-scale experiment in India." *World Development* 127: 104736.

OECD. 2019a. PISA 2018 Results (Volume I): What Students Know and Can Do, PISA, OECD Publishing, Paris.

OECD. 2019b. Education at a Glance 2019: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/f8d7880d-en.

Rockoff, J.E. 2015. “Evaluation Report on the School of One i3 Expansion." Unpublished manuscript. New York, NY: Columbia University.

Rodriguez-Segura, Daniel. "EdTech in developing countries: A review of the evidence." *The World Bank Research Observer* (2021).

Rouse, C.E. and Krueger, A.B. 2004. Putting Computerized Instruction to the Test: a Randomized Evaluation of a "Scientifically Based" Reading Program." *Economics of Education Review* 23(4): 323–338.

SanPiN. 2010. "Sanitary and Epidemiological Requirements for Conditions and Organization of Educational Process in Schools." Decree of the Chief State Sanitary Doctor of the Russian Federation, December 29, 2010. Approved 2.4.2.2821-10

Schaefer, S. and Warren, J. 2004. Teaching Computer Game Design and Construction." *Computer-Aided Design*, 36(14): 1501-1510.

Schling, Maja, and Paul Winters. 2018. "Computer-assisted instruction for child development: Evidence from an educational programme in rural Zambia." The Journal of Development Studies 54.7: 1121-1136.

Schmitt, John, and Jonathan Wadsworth. 2006. "Is There an Impact of Household Computer Ownership on Children's Educational Attainment in Britain? Economics of Education Review, 25: 659-673.

Taylor, Eric S. 2018. "New Technology and Teacher Productivity," Harvard University Working Paper.

World Bank. 2018. *World Bank Education Overview: New Technologies*, Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/731401541081357776/World-Bank-Education-Overview-New-Technologies

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. 2015. Effects of Feedback in a Computer-based Learning Environment on Students' Learning Outcomes: A Meta-analysis. *Review of Educational Research*, 85(4): 475-511.

Yanguas, M. L. 2020. "Technology and educational choices: evidence from a one-laptop-per-child program." *Economics of Education Review*. 76. Article number 101984. https://doi.org/10.1016/j.econedurev.2020.101984

Table 1: Effects of CAL Dosage 1X and Dosage 2X on Student-Reported Minutes per Week of Math and Language Homework

	(1)	(2)	(3)	(4)
	Time Math Homework		Time Language Homework	
Dosage 1X	-1.435	-1.680	-1.170	-1.066
	(1.429)	(1.499)	(1.201)	(1.253)
Dosage 2X	-0.024	-0.315	0.312	0.200
	(1.296)	(1.362)	(1.171)	(1.189)
Diff (Dosage 2X – Dosage 1X)	1.411	1.365	1.482	1.265
	(1.268)	(1.350)	(1.134)	(1.172)
Extra Covariates	NO	YES	NO	YES
Observations	5,322	5,322	5,312	5,312
R-squared	0.059	0.092	0.064	0.098
Mean Homework	43.02		44.26	

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (in math or language).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Effects of CAL Dosage 1X and Dosage 2X on Teacher-Reported Hours per Week Spent on Math and Language Class Preparation

	(1)	(2)	(3)	(4)
	Math Preparation		Language Preparation	
Dosage 1X	-0.272	-0.311	-0.099	-0.106
	(0.591)	(0.591)	(0.616)	(0.613)
Dosage 2X	-0.090	-0.121	0.196	0.114
	(0.570)	(0.557)	(0.618)	(0.605)
Diff (Dosage 2X – Dosage 1X)	0.182	0.191	0.296	0.220
	(0.555)	(0.543)	(0.590)	(0.571)
Extra Covariates	NO	YES	NO	YES
Observations	334	334	334	334
R-squared	0.179	0.199	0.255	0.277
Mean Preparation	6.33		7.17	

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (in math or language).
3) Even-numbered columns control teacher uses ICT at home, teacher ICT self-efficacy, class size.
4) Robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Effects of CAL Dosage 1X and Dosage 2X on Math and Language Test Scores

	(1)	(2)	(3)	(4)
	Math Test Score		Language Test Score	
Dosage 1X	0.108***	0.099**	0.059*	0.053
	(0.041)	(0.039)	(0.035)	(0.034)
Dosage 2X	0.101**	0.087**	-0.025	-0.015
	(0.042)	(0.039)	(0.031)	(0.031)
Diff (Dosage 2X – Dosage 1X)	-0.007	-0.012	-0.084**	-0.068*
	(0.041)	(0.039)	(0.037)	(0.036)
Extra Covariates	NO	YES	NO	YES
Observations	5,717	5,717	5,565	5,565
R-squared	0.332	0.414	0.434	0.487
Diff (Dosage 2X - 2*Dosage 1X)	-0.115	-0.111*	-0.143**	-0.121**
SE	(0.071)	(0.067)	(0.065)	(0.062)

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (baseline score in math or language). Dependent variables are standardized as z-scores (using the mean and SD of the control group).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) For the concavity test reported in the last panel, statistical significance is based on a one-tailed test.
6) *** p<0.01, ** p<0.05, * p<0.1.
7) Romano-Wolf stepdown p-values for multiple hypothesis testing calculated for the regressions (adjusting for covariates, bootstrapping 3000 times). Estimates of the effects of dosage X and dosage 2X on math scores remain statistically significant at the 1% level (p = 0.000). Estimate of the effect of dosage X on language scores statistically significant at the 5% level (instead of at the 10% level, p = 0.029), while estimate of the effect for dosage 2X on language scores remains statistically insignificant.

Table 4: Effects of CAL Dosage 1X and Dosage 2X on Student Interest in Math and Language

	(1)	(2)	(3)	(4)
	Math Interest		Language Interest	
Dosage 1X	0.086**	0.087**	0.094**	0.079**
	(0.036)	(0.036)	(0.039)	(0.038)
Dosage 2X	0.049	0.048	0.019	0.022
	(0.037)	(0.037)	(0.040)	(0.041)
Diff (Dosage 2X – Dosage 1X)	-0.038	-0.038	-0.075*	-0.057
	(0.038)	(0.038)	(0.039)	(0.040)
Extra Covariates	NO	YES	NO	YES
Observations	5,180	5,180	4,893	4,893
R-squared	0.132	0.141	0.151	0.176
Diff (Dosage 2X - 2*Dosage 1X)	-0.124**	-0.125**	-0.169***	-0.136**
SE	(0.064)	(0.064)	(0.067)	(0.067)

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (in math or language). Dependent variables are standardized as z-scores (using the mean and SD of the control group).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) For the concavity test reported in the last panel, statistical significance is based on a one-tailed test.
6) *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A1: Balance Check among Treatment Arms (Dosage 2X, Dosage 1X, and No Dosage) and Summary Statistics

	(1) Math score	(2) Language score	(3) Reading score	(4) Vocabulary score	(5) Female (1/0)	(6) Teacher ICT at home	(7) Teacher ICT self-efficacy	(8) Class size
Dosage 1X	-0.055	-0.051	-0.013	0.040	0.014	0.398	0.011	0.085
	(0.046)	(0.039)	(0.045)	(0.039)	(0.013)	(0.749)	(0.060)	(0.072)
Dosage2X	-0.091*	-0.043	0.014	-0.008	0.023	0.748	0.096*	-0.035
	(0.049)	(0.042)	(0.044)	(0.043)	(0.014)	(0.697)	(0.057)	(0.075)
Dosage 2x – Dosage 1X	-0.036	0.008	0.027	-0.048	0.009	0.350	0.084	-0.120
SE	(0.053)	(0.043)	(0.044)	(0.038)	(0.014)	(0.718)	(0.061)	(0.080)
Full Sample								
Mean	0	0	0	0	0.52	2.67	6.06	21.6
Std Dev	1	1	1	1	0.50	0.47	0.59	6.60
N	6,052	5,838	5,838	6,052	5,742	5,903	5,903	6,253
R2	0.238	0.159	0.150	0.163	0.011	0.248	0.224	0.236

Notes:

1) All regressions control for strata (block) fixed effects.
2) Cluster (school)-adjusted robust standard errors in parentheses.
3) *** p<0.01, ** p<0.05, * p<0.1.
4) Joint tests of all baseline covariates simultaneously shows no significant difference between Dosage 1X and Control (p-value: 0.265), Dosage 2X and Control (p-value: 0.178) or Dosage 2X and Dosage 1X (p-value = 0.160).

Appendix Table A2: Balance Check among Treatment Arms, Non-Missing Students

	(1) Math score	(2) Language Score	(3) Reading score	(4) Vocabulary score	(5) Female (1/0)	(6) Class size	(7) Teacher ICT at home	(8) Teacher ICT self-efficacy
Dosage 1X	-0.061	-0.050	0.017	-0.049	0.006	0.538	0.018	0.083
	(0.047)	(0.047)	(0.041)	(0.040)	(0.014)	(0.747)	(0.060)	(0.073)
Dosage 2X	-0.087*	-0.005	-0.005	-0.062	0.017	0.855	0.099*	-0.030
	(0.048)	(0.045)	(0.044)	(0.044)	(0.015)	(0.706)	(0.057)	(0.076)
Observations	5,552	5,205	5,205	5,552	5,495	5,958	5,619	5,619
R-squared	0.187	0.171	0.151	0.164	0.011	0.235	0.243	0.222
Diff (Dosage 2X - Dosage 1X)	-0.026	0.046	-0.022	-0.014	0.012	0.317	0.081	-0.113
SE	(0.051)	(0.046)	(0.039)	(0.045)	(0.014)	(0.727)	(0.061)	(0.081)

Notes:

1) All regressions control for strata (block) fixed effects.
2) Cluster (school)-adjusted robust standard errors in parentheses.
3) *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A3: Heterogeneous Effects of CAL Dosage 1X and Dosage 2X on Math and Language Test Scores, by Student Gender

	(1)	(2)	(3)	(4)
	Math Test Score		Language Test Score	
Dosage 1X	0.096*	0.076	0.065	0.043
	(0.051)	(0.053)	(0.046)	(0.055)
Dosage 2X	0.080	0.050	0.003	-0.004
	(0.051)	(0.056)	(0.041)	(0.055)
Female	0.049	-0.040	0.112***	0.143***
	(0.040)	(0.044)	(0.040)	(0.051)
Female * Dosage 1X	0.025	0.028	-0.009	-0.019
	(0.058)	(0.065)	(0.056)	(0.072)
Female * Dosage 2X	0.043	0.040	-0.051	-0.036
	(0.056)	(0.061)	(0.053)	(0.068)
Extra Covariates	NO	YES	NO	YES
Observations	5,716	5,716	5,565	5,565
R-squared	0.333	0.163	0.436	0.139

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (baseline score in math or language).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A4: Effects of CAL Dosage 1X and Dosage 2X on Interest in Math and Language, by Gender

	(1)	(2)	(3)	(4)
	Math Interest		Language Interest	
Dosage 1X	0.093*	0.091*	0.092	0.075
	(0.050)	(0.049)	(0.057)	(0.056)
Dosage 2X	0.071	0.068	0.024	0.021
	(0.053)	(0.053)	(0.060)	(0.060)
Female	0.011	-0.040	0.242***	0.275***
	(0.052)	(0.106)	(0.052)	(0.106)
Female * Dosage 1X	-0.013	-0.008	0.005	0.007
	(0.068)	(0.068)	(0.068)	(0.068)
Female * Dosage 2X	-0.043	-0.039	0.003	0.006
	(0.071)	(0.070)	(0.073)	(0.074)
Extra Covariates	NO	YES	NO	YES
Observations	5,180	5,180	4,893	4,893
R-squared	0.132	0.141	0.166	0.177

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (in math or language).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A5: Heterogeneous Effects of CAL Dosage 1X and Dosage 2X on Math and Language Test Scores, by Student Ability (Above and Below Median Baseline Score)

	(1)	(2)	(3)	(4)
	Math Test Score		Language Test Score	
Dosage 1X	0.118**	0.125**	0.020	0.027
	(0.056)	(0.051)	(0.052)	(0.050)
Dosage 2X	0.140**	0.136***	-0.083*	-0.069
	(0.055)	(0.052)	(0.048)	(0.047)
High Ability (>50%)	0.571***	0.483***	0.211***	0.139***
	(0.049)	(0.045)	(0.049)	(0.044)
High Ability * Dosage 1X	-0.042	-0.074	0.027	0.016
	(0.063)	(0.060)	(0.058)	(0.054)
High Ability * Dosage 2X	-0.086	-0.106*	0.086	0.091*
	(0.063)	(0.059)	(0.057)	(0.054)
Extra Covariates	NO	YES	NO	YES
Observations	5,552	5,552	5,205	5,205
R-squared	0.381	0.444	0.467	0.510

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (baseline score in math or language).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A6: Effects of CAL Dosage 1X and Dosage 2X on Interest in Math and Language, by Student Ability (Above and Below Baseline Median Score)

	(1)	(2)	(3)	(4)
	Math Interest		Language Interest	
Dosage 1X	0.097*	0.099*	0.086*	0.075
	(0.053)	(0.051)	(0.050)	(0.049)
Dosage 2X	0.040	0.040	-0.072	-0.065
	(0.055)	(0.054)	(0.054)	(0.054)
High Ability (>50%)	0.150***	0.115**	0.056	-0.050
	(0.043)	(0.045)	(0.047)	(0.052)
High Ability * Dosage 1X	-0.021	-0.027	0.008	0.005
	(0.063)	(0.063)	(0.065)	(0.065)
High Ability * Dosage 2X	0.021	0.015	0.136**	0.143**
	(0.067)	(0.067)	(0.067)	(0.065)
Extra Covariates	NO	YES	NO	YES
Observations	5,034	5,034	4,594	4,594
R-squared	0.136	0.141	0.155	0.177

Notes:

1) Dosage 1X is 10 items (approximately 20-25 minutes) of CAL per subject per week, Dosage 2X is 20 items (approximately 40-50 minutes) of CAL per subject per week. Left out category is pure control (no CAL).
2) All columns control for baseline counterpart of dependent variable (in math or language).
3) Even-numbered columns control for all baseline test scores (math, language, reading, and vocabulary), student gender, teacher uses ICT at home, teacher ICT self-efficacy, and class size.
4) Cluster (school-level)-robust standard errors in parentheses.
5) *** p<0.01, ** p<0.05, * p<0.1.

Appendix A: Computer-Assisted Learning Software – Example Screenshots

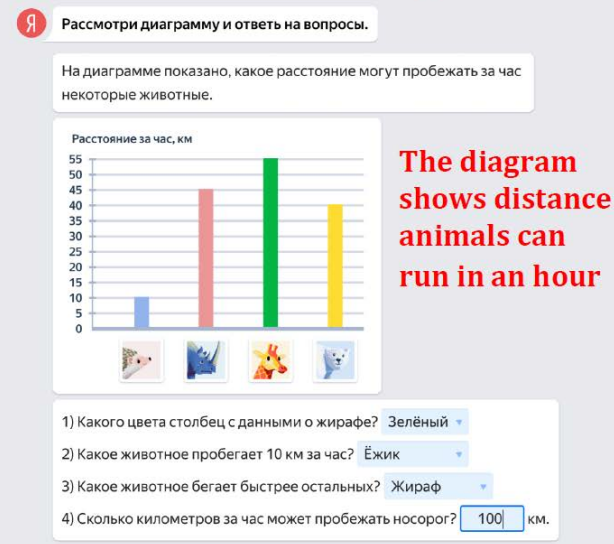
Example:
Literacy Items
(e.g. Science)

Работа
с информацией

Учат анализу и извлечению информации из диаграмм, схем, таблиц и текста.

Item purpose:
Learn to analyze and extract information from charts, diagrams, tables and text.

Item: View the chart and answer questions



The diagram shows distance animals can run in an hour

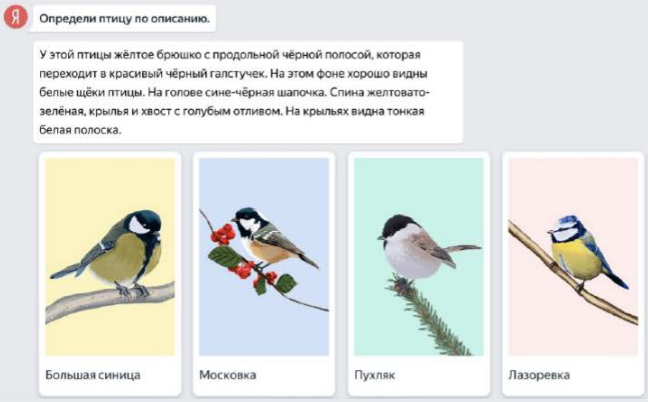
1) **What color is the giraffe data column?**
2) **Which animal runs 10km an hour?**
3) **Which animal runs faster than the others?**
4) **How many kilometers per hour can a rhino run?**

Example:
Items that make cross-subject connections

Межпредмет-
ные задания
Формируют научное мировоззрение и повышают познавательную активность.

Item Purpose:
Develop a scientific worldview and increase cognitive activity

Item: Identify the bird by its description



Learning process: scaffolded feedback for students

Дети учатся: попытки, подсказки

Помогают двигаться в «зону ближайшего развития».

Feedback helps move students to the "zone of proximal development"

Item: emphasize the subject and predicate

После первой попытки «Слабая» подсказка (не вмешиваемся, даем пробовать ещё)

After the first attempt, a "Weak" hint (do not intervene, let's try again)

После второй попытки «Сильная» подсказка (отмечаем верное, даем пробовать ещё)

After the second attempt, a "Strong" hint (emphasize the correct, let's try again)

После третьей попытки Показываем ответ, фокусируем внимание на проблемных зонах

After the third attempt, show the answer, focus on the problems

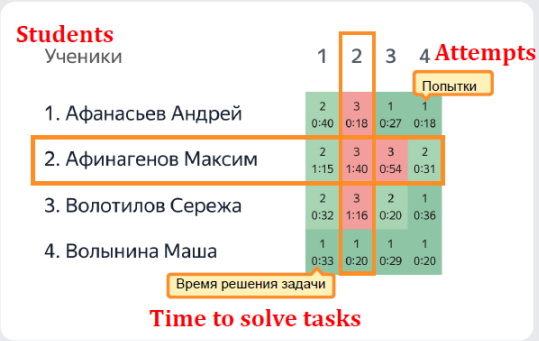
Feedback to Teachers

Результаты учеников

Считаются автоматически

Выявляют проблемные задания (по столбцам) и учеников, требующих внимания

Teachers can analyze regularly updated data to identify problems with tasks (columns) and students who require attention (rows)



Appendix B: Psychometric Properties of the Exams

The exams (collectively known as the PROGRESS toolkit) were specifically developed to assess student achievement in grade 3 in Russian schools. The exams are typically (as in our study) administrated twice: once at the beginning of the school year and once at the end of the school year.

The exams cover four areas: math, language, reading and vocabulary. Exam items for these areas were chosen based on the Russian Federal Standards for primary education. The math and language areas include 5 thematic blocks each with respectively 30 and 66 items in total. The reading and vocabulary areas include two blocks of items each with 42 and 39 items in total. The items are of different formats such as multiple choice, short answer, and matching. All items are scored dichotomously.

The language test is aimed at assessing mastery of the Russian language, its grammatical and lexical norms, as well as the ability to apply this knowledge appropriately in context. The language test consists of five blocks of items which respectively assess a student's ability to: (a) choose which of two words can best be used in a given sentence; (b) work with synonyms; (c) determine the inaccuracy of word usage; (d) choose the phraseological turn appropriate to context; (e) choose appropriate language in speech.

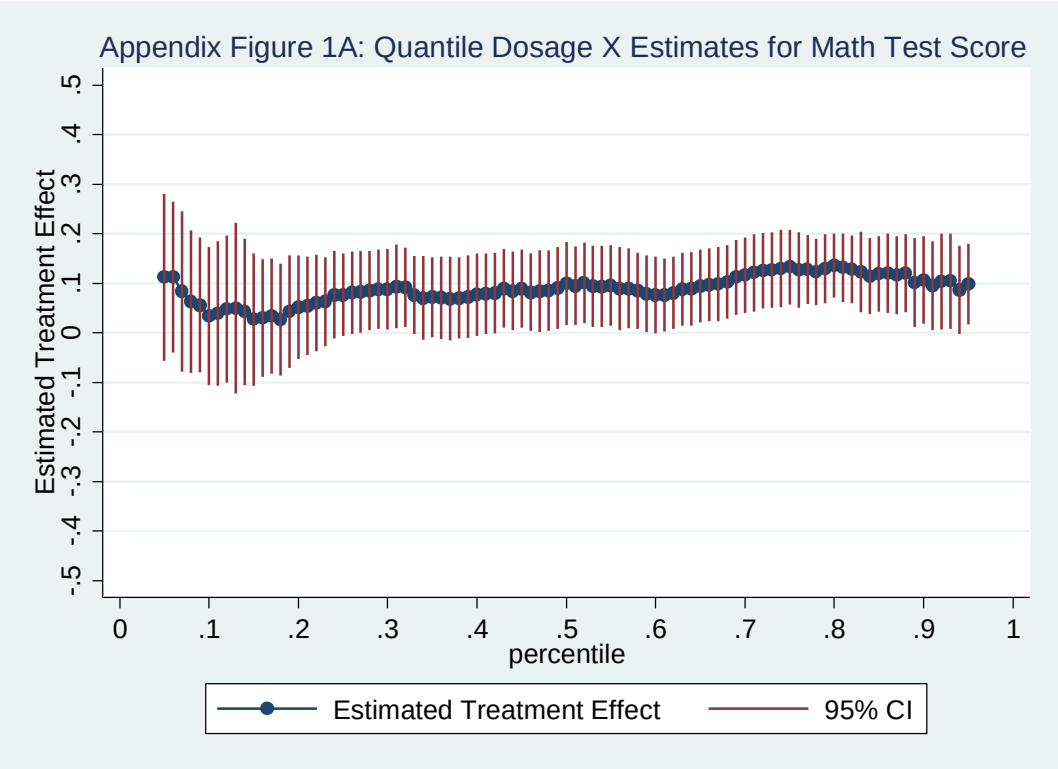
The vocabulary test is aimed at assessing a student's knowledge of basic vocabulary. It consists of two blocks of items which respectively assess the student's ability to: (a) match pictures and words; (b) match meanings and words.

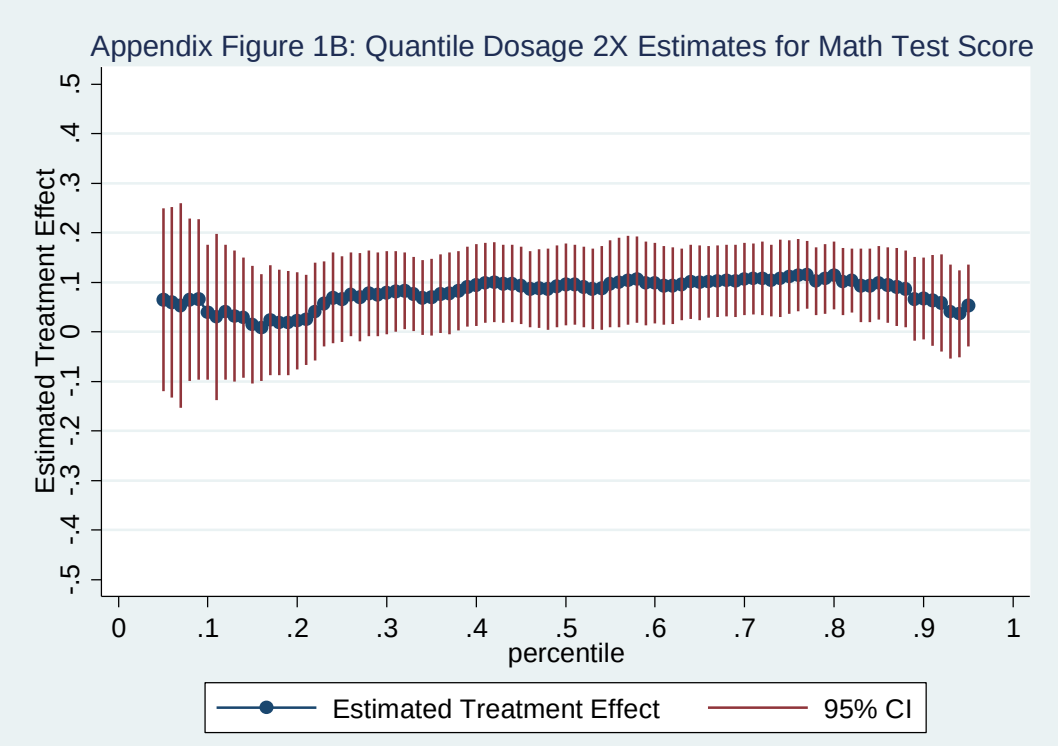
Finally, the reading test is developed on the theoretical basis of the PIRLS (Progress in International Reading Literacy Study) framework, an international assessment designed to measure reading achievement at the fourth-grade level. The reading test consists of two blocks of items in which students are asked to (a) choose words to fill-in-the-blank of a sentence; (b) reading comprehension.

Testing is conducted during two 40-minute sessions. One session tests language and reading, while another session tests math and vocabulary. Testing is computer-based. Students are assessed in schools' computer rooms under the supervision of proctors.

In our study, the exams exhibited good psychometric properties. The constructs underlying the four exam areas (math, language, reading, and vocabulary) were essentially unidimensional. All items demonstrated good model fit. We convert percent correct scores into z-scores as is standard in the economics literature. Exam reliability (Cronbach's alpha and Person reliability) varied from 0.82 to 0.96. No items demonstrated floor or ceiling effects (during the baseline or the endline). There was no evidence of differential item functioning (DIF) by gender or local region.

Appendix C: Figure 1: Quantile Effects of Dosage 1X and Dosage 2X (each versus Control) on Math Test Scores





Appendix C: Figure 2: Quantile Effects of Dosage 1X and Dosage 2X (each versus Control) on Language Test Scores