Does computer-aided instruction improve children's cognitive and non-cognitive skills?: Evidence from Cambodia

Hirotake Ito* Keiko Kasai † Hiromu Nishiuchi‡

Makiko Nakamuro§¶

September 14, 2019

Abstract

This paper examines the causal effects of computer-aided instruction (CAI) on children's cognitive and noncognitive skills. We ran a clustered randomized controlled trial at five elementary schools with 1,600 students near Phnom Penh, Cambodia for three months. The results suggest that average treatment effects on cognitive skills are positive and statistically significant, while their hours of study were not changed both at home and classroom. It indicates that CAI is successful to improve student's learning productivity. Furthermore, it is found that the CAI-based app raises students' subjective expectation to attend college in the future.

Keywords: CAI, cluster-randomized controlled trial, noncognitive skills
JEL classifications: I21, I25, I30

*Corresponding author. Graduate School of Media and Governance, Keio University

†Keio Research Institute at SFC

‡Graduate School of International Management, Yokohama City University

§Faculty of Policy Management, Keio University

¶This study is conducted as a part of the Project "Research on the Improvement in Resource Allocation and Productivity among the Healthcare and Education Service Industries" undertaken at Research Institute of Economy,Trade and Industry (RIETI). We thank supports from Kei Kawashima, Kodai Tokumaru, and Daiki Watanabe (Hanamaru-Lab, Inc. Ltd) in running the experiments in Cambodia, and Makoto Yano, Masayuki Morikawa, Kyoji Fukao and Tomohiko Inui for insightful comments. We also gratefully acknowledge the financial support received from the MEXT/JSPS KAKENHI Grant Number: 18H05314.

1 Introduction

The World Bank recently made reference to a "learning crisis" (World Bank, 2017), according to which a large proportion of students in developing countries are failing to acquire even foundational skills at school, e.g., the basic math that is required when buying and selling in markets, handling household budgets, or transacting with banks or other financial institutions.

While many low-income countries have increased primary school enrolments rapidly in recent decades, they often face substantial obstacles in avoiding a learning crisis. First, the increases in primary school enrolments have occurred along with increases in education inputs, such as teachers and other school resources. However, any decline in per capita inputs is likely to reduce the quality of primary education. Second, hiring high-quality teachers is difficult in many developing countries because they are paid less than other comparably qualified professionals, particularly in urban areas. Third, any substantial gap between the abilities of low- and high-achieving students makes it difficult for teachers to set their level of instruction appropriately. Such situations produce a mismatch between the level of a teacher's instruction and students' level of proficiency (Glewwe and Muralidharan, 2016).

New technologies offer promising ways to mitigate such problems in developing countries. Although computer access in classrooms does not improve students' learning, as shown in Barrera-Osorio and Linden (2009), well-designed computer-assisted learning (CAL) allows students to access high-quality instructional materials even in the presence of severe teacher shortages and to learn at their own pace and proficiency. However, the empirical evidence on the effect of computer-aided instruction (CAI) is mixed. In India, CAI was found to improve student performance substantially, especially for low-achieving students (Linden, 2008), while the One Laptop Per Child programs in Peru and Uruguay had no impact on student reading or math abilities (Cristia et al., 2012; De Melo, Machado and Miranda, 2014).

This study was designed to rigorously estimate the causal impact of CAI on students' cognitive and noncognitive skills, in collaboration with the government of Cambodia, the Japan International Cooperation Agency (JICA), and Hanamaru Lab, a Japanese private company that developed a personalized computer-assisted app, called Think!Think! The primary objective of Think!Think! is to develop foundational math skills for elementary school students.

To examine the effect of Think!Think!, we ran a clustered randomized controlled trial (RCT) involving 1,656 students from grade 1 (G1) to grade 4 (G4) at five public elementary schools near Phnom Penh from May to August, 2018. Because each school has two classes in each grade, students were randomly assigned to either one of the 20 treatment classes that used Think!Think! during the three-month intervention or to one of the 20 control classes.

Our results suggest that the average treatment effects on cognitive skills measured by several types of math achievement tests and IQ tests are positive and statistically significant. The effect size is large, especially compared with previous studies conducted in developing countries: the preferred point estimates on student achievements are 0.68–0.76 standard deviations and IQ scores 0.65 standard deviations even after controlling for the prior score in the baseline survey, gender, grade, birth month, parental education, and schools' time-invariant characteristics. Furthermore, the CAI-based app raises students' subjective expectations of attending college in the future. However, there is no significant effect on noncognitive skills, such as motivation and self-esteem.

The remainder of this paper proceeds as follows. Section 2 provides a literature review. Section 3 explains the research design and data. Section 4 presents empirical specifications and the main results on cognitive and noncognitive skills. Section 5 concludes and provides policy implications.

2 Literature Review

Previous studies defined investment in computers by schools as either (i) information communication technology (ICT) or (ii) CAI. In recent years, CAI programs, which do not necessarily require an Internet connection, have become more widely used in public schools. However, while several studies showed that well-designed CAI programs appear to have strong and positive effects on the math or science abilities of weaker students, especially in developing countries, other studies found insignificant effects on reading and language test scores. For example, Rouse and Krueger (2004) ran a large-scale RCT using the computer software program Fast For Word for G3 to G6 students in an urban district in northwestern United States. Their results showed that the effect of this program on language and reading skills is small and statistically significant. Banerjee et al. (2007) examined the effect of a CAI program for G4 students in urban India. The students who were randomly assigned to treatment schools increased their math achievements by 0.47 standard deviations, mainly because of improvement among the poorer-performing children. Surprisingly, this positive effect remained even after the programs were terminated, although the size of the effect decreased to about 0.10 standard deviations.

In the field of economics, investments in computers, the Internet, software, and other technologies have been analyzed in the context of an education production function. Bulman and Fairlie (2016) pointed out that the binding constraint in the model is often the amount of time available for instruction, which is regarded as one of the educational inputs. In other words, this trade-off between time spent using a computer in class and time spent on traditional instruction makes it more difficult to determine whether schools should use CAI programs or more traditional instruction. However, many studies, including Rouse and Krueger (2004) and Banerjee et al. (2007), have estimated the effect of supplemental education or remedial education with CAI programs outside of class.

To deal with these issues, Barrow, Markman and Rouse (2009) developed a trial in which middle school students in randomly assigned treatment classes were taught using CAI, while students in the control classes were taught traditionally in class. This enabled a comparison of the effects of the newly developed CAI program and more traditional instruction under limited school resources and time constraints. The two-year experiment found that the treatment students improved their math ability by at least 0.17 standard deviations more than their counterparts. Carrillo, Onofa and Ponce (2011) conducted a similar experiment in Ecuador for elementary school students. Using CAI in class, instead of traditional instruction, helped to improve math performance, but not language acquisition. However, a recent study on middle schools in urban India showed that using CAI in class has a greater impact on both math and language abilities (Muralidharan, Singh and Ganimian, 2019). Their IV estimates suggested that treatment students performed 0.37 standard deviations higher in math and 0.23 standard deviations higher in Hindi during the five-month intervention. They also found that the achievement gains were greater for academically weaker students. Our empirical analysis follows that of Muralidharan, Singh and Ganimian (2019) and tests whether CAI programs is effective for younger-aged children in relatively disadvantaged areas of the developing country.

3 Methodology and Data

3.1 Background

Our study targets five public elementary schools located within a radius of approximately 10 km around Phnom Penh. Because these schools did not receive any aid or assistance from other development agencies during the period of our intervention, we can rule out any confounding factors from other external interventions. The majority of households around the schools engage in farming and fishing to generate income. Only a small proportion of parents have tertiary education. The locations of these five schools are illustrated in Figure 1.

3.2 Baseline and Follow-up Surveys

Prior to the intervention, we conducted baseline surveys in class from May 21 to May 25, 2018 with the full cooperation of teachers and staff. The baseline survey included two sets of 40-minute achievement tests for G3 and G4 students, 40-minute IQ tests for all students, and 20-minute surveys for all students and parents.

To measure students' cognitive skills, two sets of achievement tests were used: the National Assessment Test (NAT) administered by Cambodia's Ministry of Education, Youth and Sports for G3 students and Trends in International Mathematics and Science Study (TIMSS) administered by the International Association for the Evaluation of Educational Achievement (IEA) for G4 students. We selected exams that the students in our intervention had not sat previously. As there are no standardized tests to measure the cognitive abilities of younger students, we did not administer achievement tests for the G1 and G2 students. Instead, we administered two sets of age-appropriate IQ tests in the baseline survey. One of the IQ tests—the "new Tanaka B-type intelligence test" (Tanaka, Okamoto and Tanaka, 2003)—has long been used in Japan and other countries in Asia as an age-appropriate measure of children's cognitive skills. The Tanaka B-type intelligence tests was translated into the local language and also modified appropriately for the local environment (e.g., illustrations of local banknotes, food, people, etc). The other intelligence test conducted during the baseline survey was the Goodenough Draw-a-Man (DAM) test (Goodenough, 1926). In this test, students are asked to complete drawings of a whole person(s) on a piece of paper for 10–15 minutes. Several examples of children's drawings collected during our baseline survey appear in Figure 2. Although the validity of this test as a measure of intelligence has been criticized, the literature suggests that the DAM test is effective in screening for lower levels of intelligence for 5–12-year-old children (Scott, 1981).

The survey of all G1–G4 students asked them to provide demographic information, such as gender, grade, birth month, hours of study at home, and subjective likelihood of attending college in the future. The survey also included a set of questionnaires to measure noncognitive skills, such as the Rosenberg self-esteem

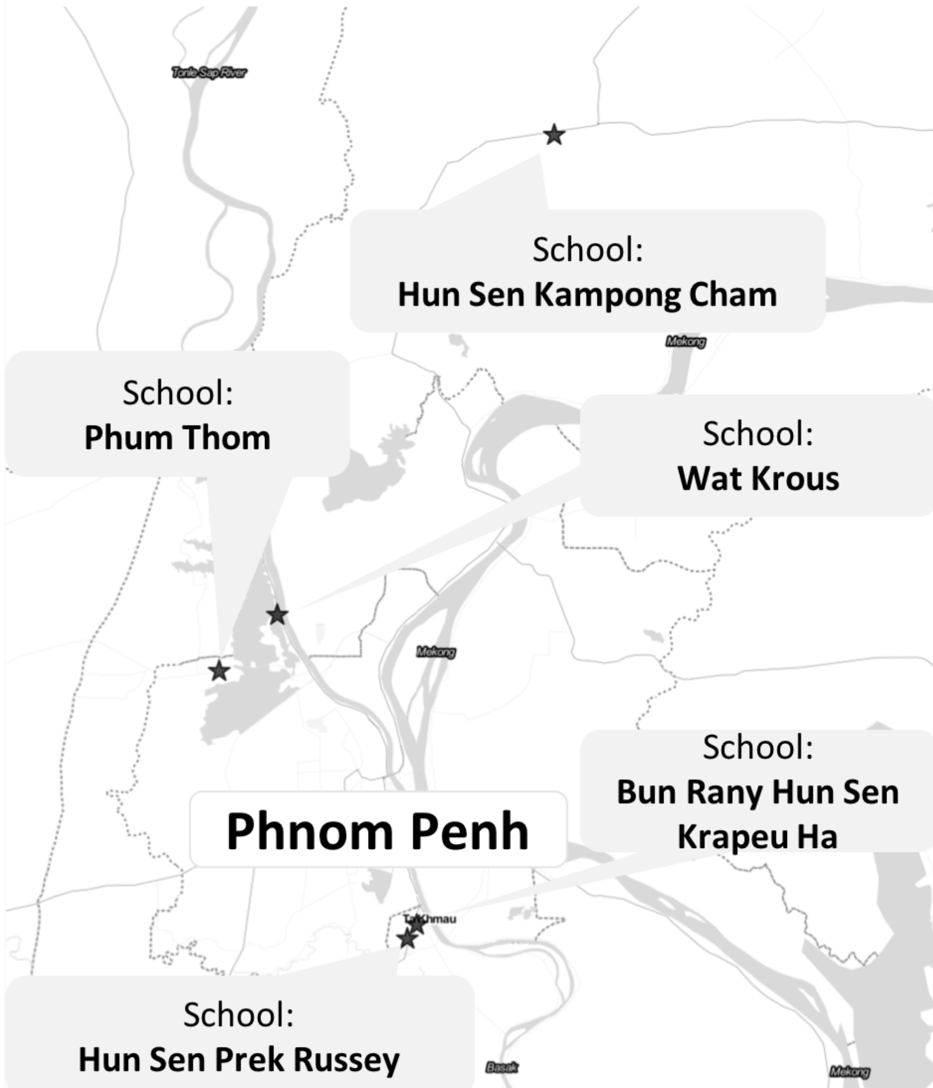

Figure 1: The location of target schools

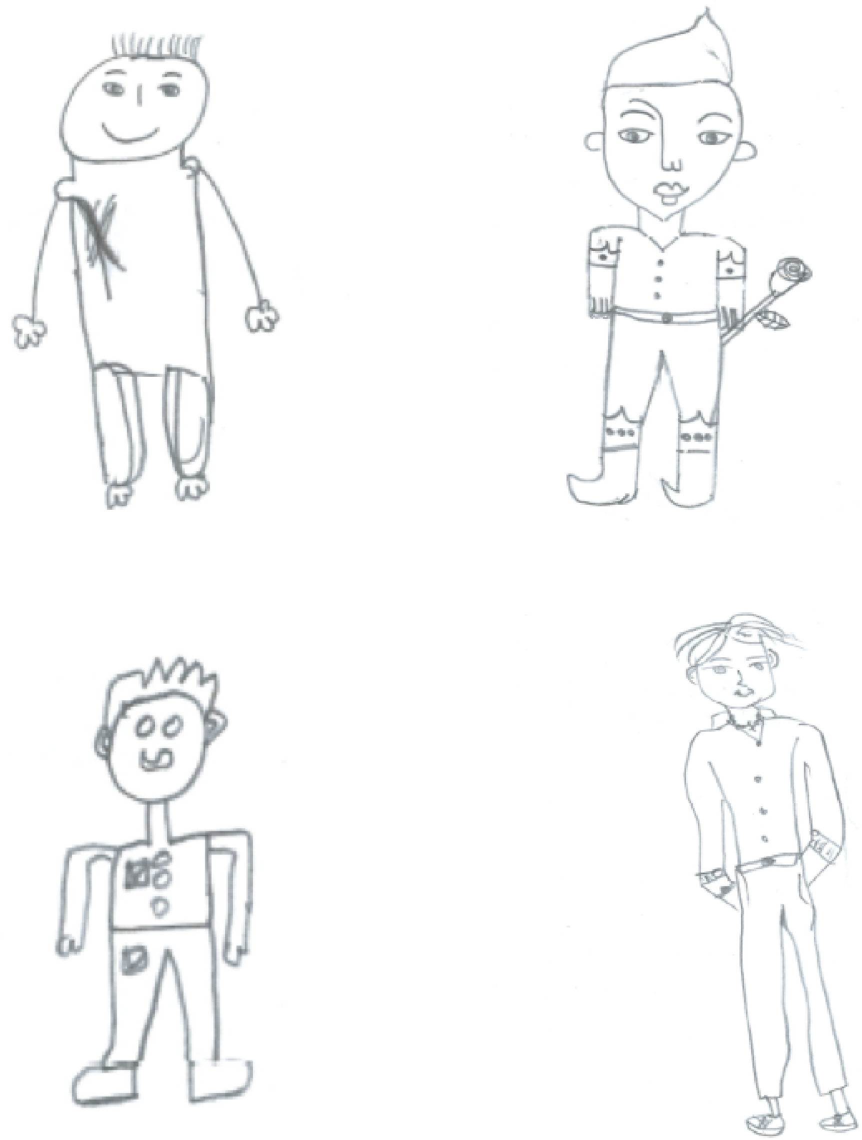

Figure 2: Samples of Draw-a-Man Test

scale (Rosenberg, 1965) and internal and external motivation for study (Sakurai and Takano, 1985). The survey of parents asked for socioeconomic status, such as parental educational backgrounds.

Following the three-month intervention, a follow-up survey was conducted from August 16 to August 25. We again administered the same sets of achievement tests, IQ tests, and questionnaires for students, focusing only on time-varying variables, such as willingness to attend college and time spent studying at home.

Out of 1656 students who officially registered to our target schools, 77.2% of them participated both in baseline and follow-up surveys, although 6.3% did only baseline. The sample attrition may be a great threat to reduce the comparability of treatment and control. If our intervention is successful, the low-achieving students assigned to the treatment group may not drop out during the intervention, while their counterpart low-achieving students assigned to control group may drop out of school altogether. In this case, the estimated impact of this intervention may be downward biased. We calculated the attrition rate for both treatment and control groups and checked whether the different characteristics of students dropped out of the two groups. Fortunately, there is no evidence of differential attrition rates and different types of attrition in the treatment and control groups. However, we still do not know much about 9.2% of students who attended neither baseline nor follow-up surveys. According to the latest World Bank Indicators, the drop out rate in Cambodia nationwide is 9.4% in 2017. Because our intervention was implemented in the last three months at the end of semester, some of them may be drop out of school before or during the intervention.

3.3 Education App: Think!Think!

The app Think!Think! used in our intervention was originally developed by Hanamaru Lab, taking full advantage of its substantial experience of operating a large number of cramming schools for school-aged children in Japan. This app is designed to develop the foundational math skills of elementary school students (Figure3). More specifically, the app incorporates adaptive learning using an original algorithm and provides math problems, materials, and instruction to reflect the proficiency level of each individual student.

Think!Think! was modified for elementary school students in Cambodia to meet local curriculum standards and was translated into the local language, Khmer. Students who were assigned to treatment classes were provided with free access to a tablet/laptop to use Think!Think! in class. CAI often requires addi-

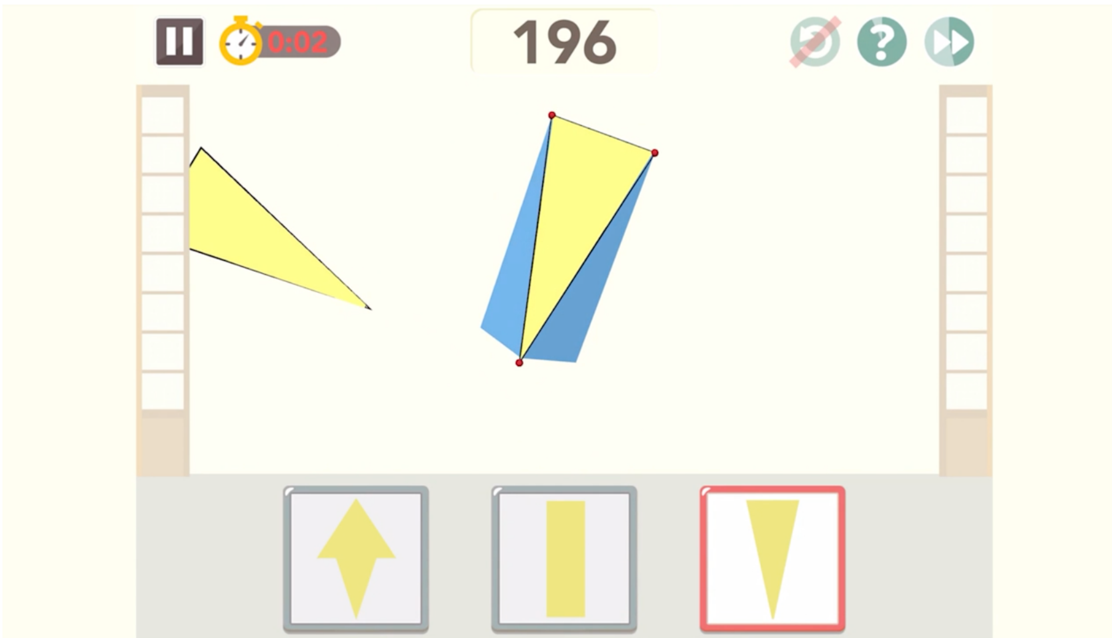

Figure 3: Sample problem

tional teaching staff in class. In our intervention, we provided three additional staff with no teaching experience to advise students on technical matters and time management.

3.4 Clustered RCT

If we were to allow students to access the CAI-based app based on their own preferences, the app would most likely be used by higher-achieving students. Students who have sought to access higher quality of education, including the exposure to new technology, are not as enthusiastic to study, on average, as those who never did. Random assignment of access to the CAI-based app avoids this selection bias.

Students in the treatment classes used Think!Think! for approximately 30 minutes each day. Peer effects are a potential threat to the internal validity of this experiment and interactions between students may violate the stable unit treatment value assumption (SUTVA). To avoid this situation, besides the fact that clustered RCT is more common in education as noted in the literature, we randomized students within intact classrooms, rather than individual students within them.[1]

Because each school has two classes in each grade, we randomly selected one of those classes as the treatment group. This created 20 treatment classes (with 840 students) and 20 control classes (with 816 students) across the five schools. However, there is still the concern that students in the treatment classes would talk to their friends in the control classes at the same school about what they had learned. To reduce the risk of such spillover, we did not allow the treatment-group students to access Think!Think! outside of class. Furthermore, they were not allowed to take their tablet/laptop home. Despite the relatively short period of intervention of three months, the students were enthusiastic about using Think!Think![2].

4 Econometric Specification and Results

4.1 Econometric Specification

To identify the causal effect of using Think!Think!, we conduct ANCOVA using the following model and identify the effect of using CAI. Our equation of interest is:

$$Y_{i,j,t} = \alpha + \beta T_{i,j,t} + \gamma Y_{i,j,t-1} + X_{i,j,t}\delta + \varepsilon_{i,j,t}$$

where $Y_{i,j,t}$ is the outcome variable of student i in school j at time t. $T_{i,j,t}$ is access to CAI and the key independent variable of interest. $X_{i,j,t}$ is a vector of control variables, while $\varepsilon_{i,j,t}$ is the idiosyncratic error term. According to McKenzie (2012), analysis of covariance (ANCOVA) is preferred for experimental designs, rather than the difference-in-difference approach, when the autocorrelation in outcome variables between the baseline and the follow-up survey is low. Because our data are only weakly autocorrelated, we apply ANCOVA for our estimation.

The crucial identifying assumption in this empirical model is that the relationship between exposure to the CAI-based app and students' unobserved ability is

[1] However, as pointed out by (Imbens, and Wooldridge, 2009), it is technically difficult to separate out the direct effect of the intervention on an individual from the indirect effect of peers on that individual.

[2] There can be unobserved correlations between the outcome of students in the same classroom and the clustered standard errors can be used to correct for such correlations. Because there are only 40 classrooms in our experiments and the calculation of the clustered standard errors requires at least 42 clusters, as suggested by Angrist and Pischke (2009), our estimation cannot be the case to apply.

orthogonal to the error term, conditional on the controls. Under this assumption, the estimate of β in equation (1) can be interpreted as the causal impact of the CAI-based app on student outcomes.

4.2 Variable Definitions

As shown in Table 1, the outcome variable of interest denoted by $Y_{i,j,t}$ is defined as follows: Table 1 presents a balance check for the baseline survey. There is no statistically significant difference in the results of the NAT between the G3 students assigned to treatment classes and those assigned to control classes, although the G4 students in the treatment classes performed much better on the TIMSS than those in the control classes.

Another outcome variable is IQ test scores: the results of the Tanaka B-type IQ test and the DAM test are converted to mental age (MA) and the IQ scores are then calculated as MA divided by chronological age (CA) multiplied by 100. According to the descriptive statistics, the mean of the Tanaka B-type IQ test score is 78.612 with a standard deviation of 13.451 and the mean of the DAM type IQ score is 0.692 with a standard deviation of 0.207. There is no statistically significant difference between the Tanaka B-type IQ test score and the DAM score.

The next set of outcome variables, measures of noncognitive skills, are coded as the mean of a set of questionnaires specific to self-esteem and motivation. The self-esteem measure is slightly higher for the treatment students, while the motivation measure is similar across the two groups of students. All cognitive and noncognitive outcome measures are normalized a mean of zero and a standard deviation of one when we run regression analysis.

Willingness to attend college is measured on a three-point scale (from 1 = not likely to 3 = very likely) based on students' subjective expectations. Hours spent studying at home are measured on a six-point scale (from 1 = not at all to 6 = more than 4 hours). We set the minimum of this variable equal to zero and the maximum equal to four and then took the median value for categories between two (= less than 30 minutes) and five (= 2–3 hours). The key independent variable of interest denoted by $T_{i,j,t}$ is a dummy variable coded as one if students are assigned to a treatment class, and zero otherwise.

The demographic variables denoted by $X_{i,j,t}$, such as gender, age, and parental educational backgrounds, are very similar between the treatment and control students. The variable on parental education represents the highest level of education of either one of the parents. Note that this information is retrieved from the parental survey conducted at the same time as the student survey. However, unlike the 100 percent response rate of the student survey administered during class, the response rate of the parental survey was approximately 85%.

Although the observable characteristics are similar between the two groups, several outcome variables, such as the achievement score for G4 students, DAM type IQ score, and self-esteem scale, are not comparable in the baseline survey.

Because heterogeneity across groups can occur by chance even when randomization is implemented correctly and the chance of achieving homogeneity when we randomize at the group level increases with sample size, we are not concerned by heterogeneity in four of the 15 variables. However, although schools randomize the change in class composition annually, heterogeneity between the treatment and control groups may still exist because of dropout or absence on the day of the baseline survey. We thus control using the demographic variables we use for the heterogeneity check to enable a "pure" comparison.

The average treatment effect may depend on the interests of particular subgroups of students. For example, if boys are more familiar with computer-related equipment, the effect may be stronger for boys than for girls. This kind of heterogeneous effect is important for policy makers in designing policy to reflect the needs of particular subgroups. We will discuss this point in Section 4.

4.3 Results

Effect on Cognitive Skills

We start by estimating the effect of CAI on student achievement. The OLS estimates are reported in Table 2 along with heteroskedasticity-robust standard errors. Our primary focus is the estimated effect of access to Think!Think! on the NAT for G3 and on the TIMSS for G4 in the first row of the table.

Model 1 provides unconditional ANCOVA estimates. Model 2 controls only for the prior achievement score in the baseline survey. Model 3 controls for the basic demographic controls, such as gender, grade, birth month, parental education, and school time-invariant fixed effects, in addition to the prior test score.

The results clearly show that the estimated coefficients on the standardized test scores are positive and statistically significant at the 0.1 percent level (Table 2, NAT). The estimated coefficients for the sample of G3 students indicate that exogenous exposure to the CAI app raises average test scores by about 0.75 standard deviations in Model 2.

Adding demographic controls to Model 2 neither changes the magnitude of the coefficients across specifications nor improves the precision of our estimates

	ALL	Treatment (A)	Control (B)	Difference (A)-(B)
Achievement Test (NAT, G3)	0.538 (0.207, 356)	0.522 (0.198, 177)	0.554 (0.214, 179)	0.032
Achievement Test (TIMSS, G4)	0.292 (0.203, 347)	0.330 (0.187, 174)	0.252 (0.211, 173)	-0.078***
IQ Test (Tanaka-B)	78.612 (13.451, 1385)	78.432 (13.131, 700)	78.795 (13.777, 685)	0.363
IQ Test (Draw-a-man)	0.692 (0.207, 1217)	0.678 (0.206, 594)	0.705 (0.207, 623)	0.027**
Self-esteem	2.762 (0.549, 1150)	2.726 (0.596, 535)	2.794 (0.502, 615)	0.068**
Motivation	0.656 (0.142, 996)	0.652 (0.150, 471)	0.660 (0.133, 525)	0.008
Willingness to go to college	2.410 (0.771, 1051)	2.342 (0.809, 482)	2.467 (0.734, 569)	0.125***
Minutes of studying at home week	114.858 (121.313, 1511)	112.384 (115.812, 711)	117.056 (126.032, 800)	4.672
Gender (male=1, woman=0)	0.525 (0.500, 1643)	0.530 (0.499, 813)	0.519 (0.500, 830)	-0.011
Age	8.485 (1.553, 1620)	8.501 (1.573, 803)	8.470 (1.535, 817)	-0.031
Highest parental education				
College or Graduate school	0.017 (0.129, 1660)	0.012 (0.110, 818)	0.021 (0.145, 842)	0.009
High school	0.340 (0.474, 1660)	0.353 (0.478, 818)	0.328 (0.470, 842)	-0.026
Junior high school	0.222 (0.416, 1660)	0.218 (0.413, 818)	0.227 (0.419, 842)	0.009
Elementary school	0.164 (0.370, 1660)	0.160 (0.367, 818)	0.167 (0.374, 842)	0.007
no education(ref)	0.001 (0.035, 1660)	0.002 (0.049, 818)	0.000 (0.000, 842)	-0.002
Birth of Month				
Jan. - Mar.	0.228 (0.420, 1660)	0.219 (0.414, 818)	0.238 (0.426, 842)	0.019
Apr.-Jun.	0.240 (0.427, 1660)	0.258 (0.438, 818)	0.223 (0.417, 842)	-0.035*
Jur.-Sep.	0.243 (0.429, 1660)	0.251 (0.434, 818)	0.236 (0.425, 842)	-0.014
Oct.-Dec.	0.264 (0.441, 1660)	0.254 (0.436, 818)	0.273 (0.446, 842)	0.019

Note: The unit of observation is students. Means are reported in each cell along with standard deviations and the number of observations in parentheses (in this order). The column ” Difference” shows the difference between mean in treatment class and mean in control class, and the statistical significance. ” *** ” , ” ** ” , and ” * ” represent 0.1 percent, 1 percent, and 5 percent significance level, respectively.

Table 1: Descriptive statistics and balance test

in explaining the variation in test scores. Once we include the interaction term and test for heterogeneous effects for gender, grade, and parental education, we obtain small point estimates on nearly all the interaction terms, and the differences between these coefficients do not support the hypothesis of significant heterogeneous effects on test scores. Furthermore, the achievement gains are homogeneous for academically weaker students. These results are available upon request from the authors.

The results are consistent with our expectations for the G4 sample (Table 2, TIMSS). Access to the CAI app improves standardized test scores by 0.66 standard deviations per three-month exposure in Model 2. Adding controls increases the point estimates and decreases the standard errors of these estimates. At the same time, we do not find any significant heterogeneous effects of gender, grade, parental education, or initial achievement on test scores.

For all tests, our results indicate a stronger effect of CAI learning than the results of Muralidharan, Singh and Ganimian (2019), who reported estimates of 0.37 standard deviations for middle schools students over five-month interventions in urban India. Although Muralidharan, Singh and Ganimian (2019) found heterogeneous effects for academically weaker students, our results find similar achievement gains for all students.

In Table 3, the estimated coefficient on the Tanaka B-type IQ score is positive and statistically significant at the 0.1 percent level. Table 3 shows that the effect on the IQ score from Model 2 is 0.68 standard deviations. The estimated coefficient is unchanged after controlling for demographic characteristics in Model 3. However, the coefficients of the DAM score are not statistically significant, regardless of the model specification. Overall, our results indicate that the magnitude in cognitive skills appears to be very large, as compared with evidence suggested by previous literature that intervened for at least one year.

Using kernel density estimation, we obtain the probability density function for both the achievement test scores and IQ scores to compare the score distributions after the three-month intervention (Figures 4 – 6). Although the difference in the DAM score for the entire sample and even the interaction term with grade are not statistically significant, the skills of younger students seem to improve.

Effect on Noncognitive Skills and Inputs for Study

We then repeated the above approach using a set of noncognitive skills as outcomes. Unlike the results for cognitive skills, we do not find any significant effect for noncognitive skills, measured by motivation and self-esteem (Table 4). However, it is clear that the estimated coefficient on willingness to attend college is positive and statistically significant at the 0.1 percent level, indicating that students who used the CAI app during the class are more likely to believe that they would undertake more advanced education in the future. The coefficient remains constant after controlling for demographic characteristics in Model 2, which suggests that heterogeneous effects in terms of gender, grade and parental education do not exist. Although the results do not indicate a positive effect of the CAI-based app on noncognitive skills, the estimated probability density function(Figure 7 – 8) suggests the sign of slight change in younger grade.

We also estimated the effect on time spent studying at home, which is considered as the important input of education production function. As already mentioned the above, students were not allowed to bring the tablet-PC to their own homes. It is thus convincing that we do not find any significant effect to make study longer at home. However, students in treatment classes sharply raised their achievements, although their hours of study was not changed both at home and classroom. It indicates that CAI is successful to improve student's learning efficiency and productivity.

Dependent Variable	NAT			TIMSS		
	Model1	Model2	Model3	Model1	Model2	Model3
Treatment	0.814*** (0.291)	0.745*** (0.203)	0.762*** (0.221)	0.522*** (0.135)	0.662*** (0.085)	0.683*** (0.095)
Baseline Score		✓	✓		✓	✓
Control			✓			✓
Observations	369	336	288	350	325	286
Adjusted R2	0.131	0.662	0.703	0.051	0.105	0.226

Note: The coefficients for treatment are reported. The unit of observations is student. Columns labeled as Model 1-3 show OLS estimates. Standard errors are in parentheses and clustered by the school. " *** " , " ** " , and " * " represent 0.1 percent, 1 percent, and 5 percent significance level, respectively.

Table 2: Effect of treatment: cognitive skills

Dependent Variable	IQ			Draw A Man		
	Model1	Model2	Model3	Model1	Model2	Model3
Treatment	0.705*** (0.143)	0.680*** (0.123)	0.657*** (0.115)	0.071 (0.081)	0.008 (0.088)	-0.006 (0.089)
Baseline Score		✓	✓		✓	✓
Control			✓			✓
Observations	1404	1281	1087	1390	1128	953
Adjusted R2	0.076	0.41	0.489	0.001	0.228	0.316

Note: The coefficients for treatment are reported. The unit of observations is student. Columns labeled as Model 1-3 show OLS estimates. Standard errors are in parentheses and clustered by the school. " *** " , " ** " , and " * " represent 0.1 percent, 1 percent, and 5 percent significance level, respectively.

Table 3: Effect of treatment: IQ

Dependent Variable	Motivation		Self esteem	
	Model1	Model2	Model1	Model2
Treatment	-0.017 (0.071)	0.019 (0.087)	0.011 (0.035)	0.02 (0.05)
Baseline Score	✓	✓	✓	✓
Control		✓		✓
Observations	902	758	1047	885
Adjusted R2	0.277	0.359	0.024	0.110

Note: The coefficients for treatment are reported. The unit of observations is student. Columns labeled as Model 1-2 show OLS estimates. Standard errors are in parentheses and clustered by the school. " *** " , " ** " , and " * " represent 0.1 percent, 1 percent, and 5 percent significance level, respectively.

Table 4: Effect of treatment: non-cognitive skills

Dependent Variable	Study time (in minutes)		Willingess to go to college	
	Model1	Model2	Model1	Model2
Treatment	-0.084 (0.095)	-0.093 (0.104)	0.163* (0.096)	0.174** (0.079)
Baseline Score	✓	✓	✓	✓
Control		✓		✓
Observations	1511	1145	946	804
Adjusted R2	0.037	0.028	0.038	0.027

Note: The coefficients for treatment are reported. The unit of observations is student. Columns labeled as Model 1-2 show OLS estimates. Standard errors are in parentheses and clustered by the school. ” *** ” , ” ** ” , and ” * ” represent 0.1 percent, 1 percent, and 5 percent significance level, respectively.

Table 5: Effect of treatment: study input

5 Conclusion

We examined the causal effect of CAI on children’s cognitive and noncognitive skills. In collaboration with the government of Cambodia, we ran a clustered RCT at five elementary schools around Phnom Penh over a period of three months. Students were randomly assigned to either one of 20 treatment classes that were allowed to use the CAI app instead of regular math classes during the intervention or to one of 20 control classes. Our empirical results suggest that the average treatment effect on cognitive skills measured by several types of math achievement tests and IQ tests is positive and statistically significant. The effect size is large, especially compared with those in previous studies for developing countries: the estimated coefficients on student achievement are 0.68–0.76 standard deviations and IQ scores are 0.65 standard deviations even after controlling for demographic factors. Furthermore, we found that the CAI-based app can raise students’ subjective expectation of attending college in the future. However, there is no significant effect on noncognitive skills, such as motivation and self-esteem.

Because this clustered RCT only ran over three months, whether these effects remain in the longer term requires further investigation. Although, we have to pay attention about Hawthorn effect and John-Henry effect and other possible effects carefully. Nevertheless, our results suggest that the app Think!Think! has tremendous potential to improve students’ math scores in both the short term and possibly the longer term.

References

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying education: Evidence from two randomized experiments in India." *The Quarterly Journal of Economics*, 122(3): 1235–1264.

Barrera-Osorio, Felipe, and Leigh L Linden. 2009. "The use and misuse of computers in education: Evidence from a randomized controlled trial of a language arts program." *Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL). www. leighlinden. com/Barrera-Linden*, 20.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse. 2009. "Technology's edge: The educational benefits of computer-aided instruction." *American Economic Journal: Economic Policy*, 1(1): 52–74.

Bulman, George, and Robert W Fairlie. 2016. "Technology and education: Computers, software, and the internet." In *Handbook of the Economics of Education*. Vol. 5, 239–280. Elsevier.

Carrillo, Paul E, Mercedes Onofa, and Juan Ponce. 2011. "Information technology and student achievement: Evidence from a randomized experiment in Ecuador."

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2012. "Technology and child development: Evidence from the one laptop per child program."

De Melo, Gioia, Alina Machado, and Alfonso Miranda. 2014. "The impact of a one laptop per child program on learning: Evidence from Uruguay."

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications." In *Handbook of the Economics of Education*. Vol. 5, 653–743. Elsevier.

Goodenough, Florence Laura. 1926. "Measurement of intelligence by drawings."

Imbens, Guido W., , and Jeffrey M. Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." 47: 5–86.

Linden, Leigh L. 2008. *Complement or substitute?: The effect of technology on student achievement in India.* InfoDev Working Paper, Columbia University.

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

Rosenberg, Morris. 1965. *Society and the adolescent self-image.* Princeton university press.

Rouse, Cecilia Elena, and Alan B Krueger. 2004. "Putting computerized instruction to the test: a randomized evaluation of a "scientifically based" reading program." *Economics of Education Review*, 23(4): 323–338.

Sakurai, Shigeo, and Seijun Takano. 1985. "A new self-report scale of intrinsic versus extrinsic motivation toward: learning for children." *Tsukuba psychological research*, 7: 43–54.

Scott, Linda. H. 1981. "Measuring intelligence with the Goodenough-Harris drawing test." 89: 483.

Tanaka, Kanichi, Kenroku Okamoto, and Hidehiko Tanaka. 2003. *The New Tanaka B Intelligence Scale.* Kaneko shobo.

World Bank. 2017. *World Development Report 2018: Learning to Realize Education's Promise.* The World Bank.

A Effect of Treatment: Estimated PDF function

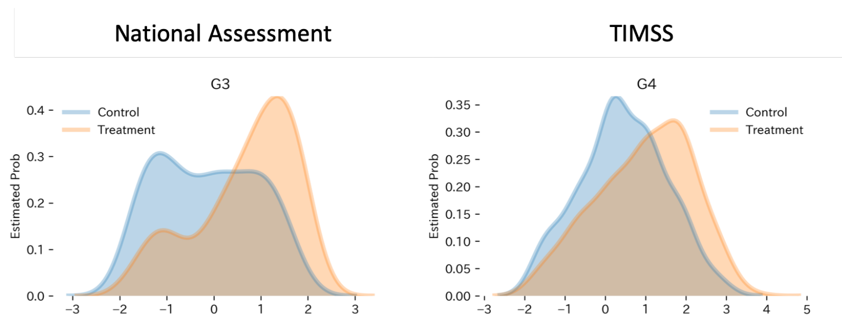

Figure 4: National assessment score and TIMSS

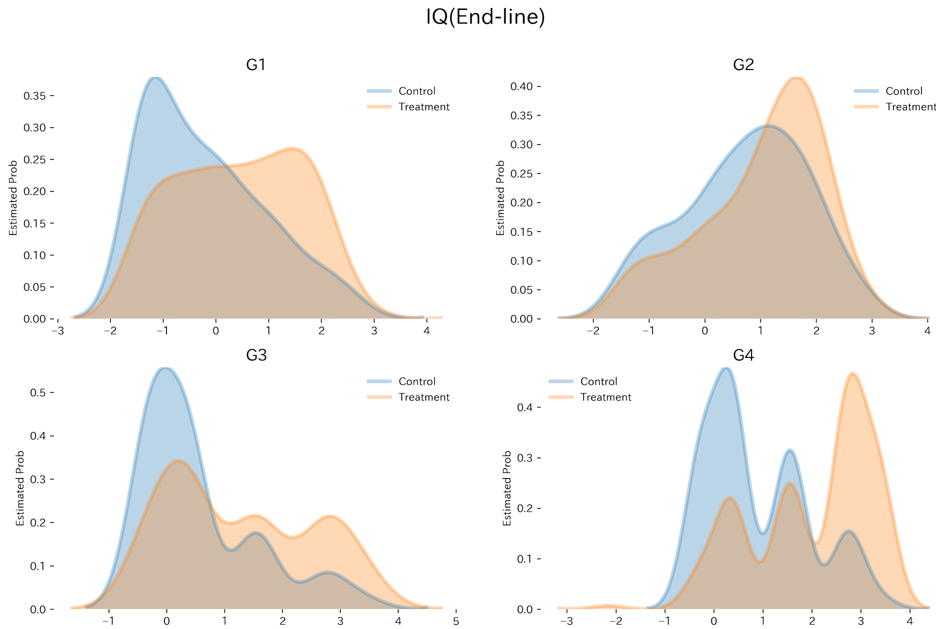

Figure 5: IQ

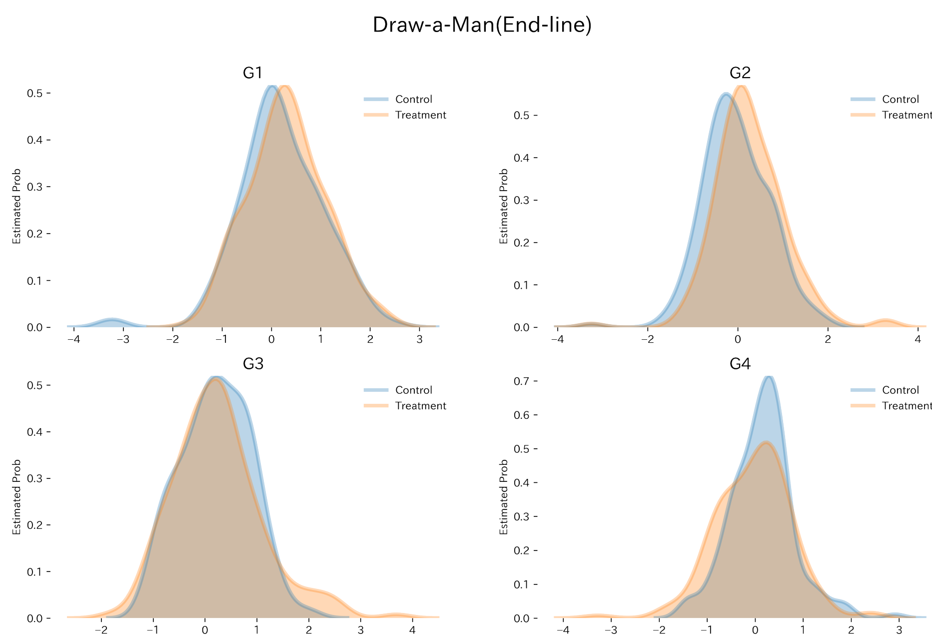

Figure 6: Draw a man test

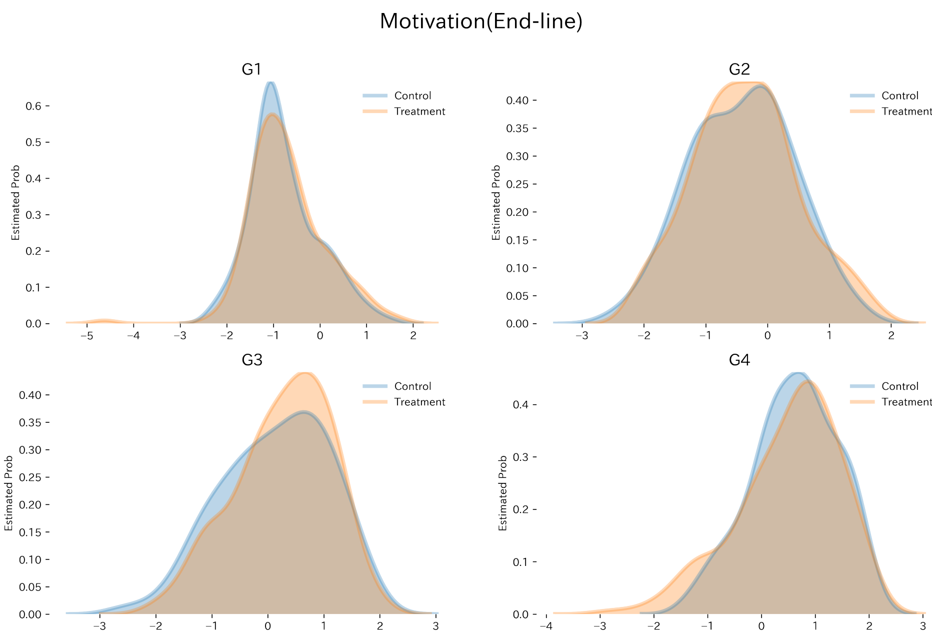

Figure 7: Motivation

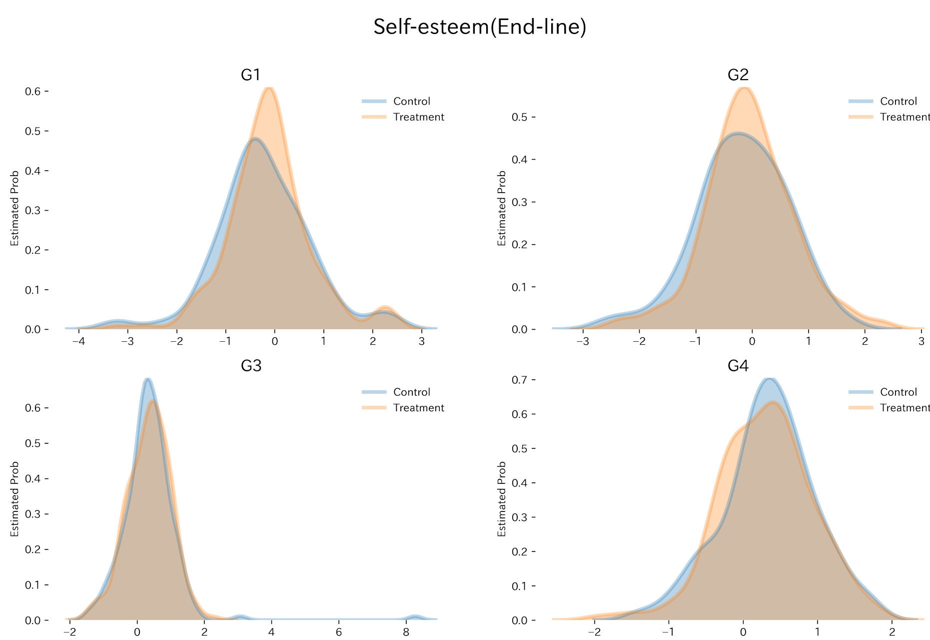

Figure 8: Self-Esteem