RIETI Discussion Paper Series 20-E-074

The Effect of Computer Assisted Learning on Children's Cognitive and Noncognitive Skills: Evidence from a Randomized Experiment in Cambodia

NAKAMURO, Makiko
Keio University

ITO, Hirotake
Keio University

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 20-E-074
September 2020

The Effect of Computer Assisted Learning on Children's Cognitive and Noncognitive Skills: Evidence from a Randomized Experiment in Cambodia[1]

Makiko Nakamuro
Hirotake Ito

Abstract

This paper examines the causal effects of computer-assisted learning on children's cognitive and noncognitive skills. We ran school-by-grade-level clustered randomized controlled trials at five public elementary schools in Cambodia. After confirming that the IQ scores of treated students significantly improved over just three months, we randomly reassigned those students either into treatment or control groups for an additional seven-month comparison. We find that students retain their cognitive skills during the additional seven-month treatment, but the initial gain diminishes for students who leave the program. Conversely, a meaningful effect on noncognitive skills is not detected immediately after the first three-month short-run program, but the effect appears to become significant and persists in the longer run.

Key words: clustered randomized controlled trial, computer-assisted learning (CAL), noncognitive skills
JEL classifications: I21, I25, I30

[1] This study is conducted as a part of the Project, "Establishing Evidence-Based Policy Making in Japan", undertaken at RIETI. The authors are grateful for helpful comments and suggestions by Kazushi Takahashi (National Graduate Institute for Policy Studies), Masahiro Shoji (University of Tokyo), and participants of RIETI seminars and the Japan Economic Association. We also thank Wonder Lab, especially Kei Kawashima, Kodai Tokumaru, Songdon Kim and Daiki Watanabe, the Japan International Cooperation Agency, and the Ministry of Education, Youth and Sport in Cambodia, for their support in implementing our experiments. We also gratefully acknowledge the financial support received from the MEXT/JSPS KAKENHI Grant Number: 18H05314. All remaining errors are our own. Keio University Shonan Fujisawa Campus IRB approval no: 238.

Introduction

The World Bank (2017) recently referred to a "learning crisis" because a large proportion of students in developing countries are failing to acquire even foundational skills at school, such as the basic math that is required when buying and selling in markets, handling household budgets, or transacting with banks or other financial institutions. The school-aged population and the number of primary school enrolments have increased rapidly in many low-income countries in recent decades.

The Kingdom of Cambodia is one country that appears to be affected by the learning crisis, as reflected in declines in per capita educational inputs under severe budget and resource constraints (OECD, 2018). In particular, hiring high quality teachers is difficult in Cambodia. Qualified teachers find the conditions in local public schools unattractive, particularly the low levels of compensation, insufficient training opportunities, and double shifts with morning and afternoon schools (UNESCO, 2006). Indeed, teacher absenteeism is significant, reaching a rate of 8.1% in primary education (OECD, 2018).

Moreover, teachers in Cambodia struggle to handle the large class sizes, producing a mismatch between the teacher's level of instruction and the students' levels of proficiency. In other words, both the quantity and the quality of teachers is an issue. Consequently, policy-makers in Cambodia began to pay considerable attention to digital personalized learning platforms, referred to as computer-assisted learning (CAL). CAL may be a promising approach that compensates for the shortage of teachers and the poor quality of teaching because it allows students to access highly individualized learning activities and to make progress at their own pace and proficiency level.

Although it seems promising in theory, the existing evidence on CAL is mixed for a comprehensive review, see Escueta et al. (forthcoming). See also Escueta et al. (2017), J-PAL (2019), and Muralidharan et al.'s (2019) online Appendix C. Although some studies have reported positive and statistically significant effects (Roschelle et al., 2010; Mo et al., 2014, 2015), others have found little or no effects on student achievements (Angrist and Lavy, 2002; Rouse and Krueger, 2004; Leuven et al., 2007; Barrera-Osorio and Linden, 2009; Malamud and Pop-Eleches, 2011; Beuermann et al., 2015;

Cristia et al., 2017; Schling and Winters, 2018). Moreover, previous studies have found heterogenous treatment effects based on student gender, grade, subject, types of school and/or parental socioeconomic status (Barrow et al., 2009; Campuzano et al., 2009; Machin et al., 2007 Shapley et al., 2009; Carrillo et al., 2011; Lai et al., 2013; Piper et al., 2016).

An important implication from the prior studies is that simply providing free laptops to K-12 students does not improve student outcomes. However, more recent experiments in India that subsidized not only hardware but also educational software have exhibited remarkable success. Specifically, Banerjee et al. (2007) and Muralidharan et al. (2019)—the two papers that are most closely related to ours—found that CAL boosted math test scores by 0.47 and 0.60 standard deviations (SD), respectively. They suggested that educational software has the potential to overcome traditional classroom constraints by replacing one-size-fits-all teaching with computer-based personalized learning, resulting in an increase in teachers' productivity per unit of teaching hour. This is consistent with a large number of recent studies showing that "Teach at a Right Level" (TaRL) is highly effective in improving students' cognitive skills (e.g., Duflo et al. 2011).

A key difference between the existing literature and our study is that we implemented our experiment during regular class hours, i.e., through an in-class program. By contrast, the prior experiments involved out-of-school remedial education that was supplementary to the regular classroom instruction (i.e., out-of-class program). In the previous literature, the research question was whether the technology could have a positive effect on test scores in *the absence of an instructional time constraint*. The most natural counterfactual of the out-of-class intervention was the student's self-study (or lack of it) at home. In the case of in-class interventions, the counterfactual was traditional teacher-led instruction in a business-as-usual curriculum. The effect of an in-class program may be smaller than that of an out-of-class program, especially when regular teaching is effective.

A few papers have attempted to compare the test score gains between in-class and out-of-class programs (Linden, 2008; Barrow et al., 2009). In particular, Linden (2008) compared two randomized experiments and concluded that in-class intervention had large and negative effects, whereas out-of-school intervention had positive effects. Such findings open up the question of whether it is worthwhile substituting traditional classroom

instructions with in-class CAL, which we believe that our paper can complement previous literature. Given that the amount of time available for instruction at school is constrained, whether schools choose the optimal level of technology relative to teacher-led instruction may be a more relevant question for policy circles.

Furthermore, we contribute to the literature by examining the effect of CAL not only on cognitive but also on noncognitive skills, measured in terms of motivation and self-esteem. In contrast, the prior literature has focused solely on test score gains. In addition, by exploiting longitudinal data, we examine short- and longer-run impacts of CAL, as much less is known about whether the initial gains remain after programs are terminated. To the best of our knowledge, this is the first experimental study to evaluate the impact of CAL on students' learning in Cambodia. In the context of the COVID-19 pandemic, the importance of distance learning through CAL is growing worldwide, particularly in developing countries. Thus, it is essential to analyze the effect of CAL in different contexts for developing countries and establish more reliable evidence of its external validity.

To examine the causal effect of in-class CAL programs on students' outcomes, we ran a clustered randomized controlled trial (RCT) involving 1,657 students from grade one to grade four at five public elementary schools near Phnom Penh. Each school has at least two classes in each grade. Students were randomly assigned to either one of the 20 treatment classes or one of the 20 control classes. For three months, treated students were allowed to access CAL during their regular math tuition, which was based on business-as-usual curriculum.

When the new academic year commenced, students were reassigned either treatment or control classes for an additional seven-month comparison. As a result, we had three groups for comparison purposes: (i) students who were assigned to treatment classes for the entire 10 months (the initial three months plus the extra seven); (ii) students who were assigned to treatment classes for the first three months and then to control group for the remaining seven months; and (iii) students who were assigned to control groups for the entire 10 months. Because in-class setting has greater adherence to our intervention, the attendance rate among treatment students during the intervention was very high. The intent-to-treat estimates are less likely to be affected by the potential self-selection of students or parents than other types

of interventions, allowing us to interpret the intent-to-treat effect as the average treatment effect.

We report four main set of results based on our experiments. Most importantly, the average treatment effects on cognitive skills are positive and statistically significant. The effect size obtained from our intervention is equivalent to, or even larger than, those of successful interventions in the existing literature. Our preferred point estimates for IQ are 0.552 SD in the short run and 0.699 SD in the longer run. Second, we find that improved cognitive skills among students are most likely attributable to the increased learning productivity per unit of hour, not to increased hours of instruction at school or studying at home. Third, the initial academic gain sharply drops by 0.085 SD and is statistically insignificant, and the program effect does not last beyond the years during which students are exposed. Fourth, the meaningful effect on noncognitive skills is not detected immediately after the first three-month short-run program. However, the effect appears to become significant in the longer run and persists even after students leave the program.

The remainder of this paper proceeds as follows. Section II explains the intervention and the experimental design. Section III describes the data and summary statistics. Section IV presents our empirical specifications and the main results. Section V discusses the mechanisms, costs, and policy implications. Section VI concludes.

I. Intervention and Experimental Design

Sample

Our study covered five public elementary schools located within a radius of approximately 10 km around Phnom Penh. The names of the schools are Phum Thom, Wat Krous, Prek Russey, Kroper Ha, and New Generation School Anuk Wat.[2] All of the selected schools are located in the countryside. The average class size of the five schools is 41.4 students, compared with the nationwide average of 45 (OECD, 2018). Teachers teach

[2] One of the five schools is a so-called "new generation school" (NGS) in which the Cambodian Government made intensive investments to improve the quality of education. There exist 172 primary schools nationwide in Cambodia. We did not find any relevant NGS-specific heterogeneity across any specifications presented in this paper.

all subjects to students of the same cohort and often engage in a traditional instructor-centered pedagogical approach based on official textbooks.

The target schools were chosen by the Cambodian Ministry of Education, Youth, and Sport, based on two criteria: (i) the school had at least two classes in each grade, allowing us to run a class-level clustered RCT, and (ii) the school did not receive any aid or assistance from other development agencies during the period of our intervention, allowing us to rule out any confounding factors from external interventions.

Our baseline survey provides information on the parents of students at the school. It indicates that 78.1% of parents had completed at least upper secondary education, and 21.9% had completed primary education. As a comparison, the official statistics for Cambodia for 2017 indicate that 62.7% of 25--34 years have completed at least secondary education and 26.0% have completed primary education (National Institute of Statistics, Ministry of Planning, 2018). Although it is difficult to make a simple comparison, our sample may be drawn from slightly higher socioeconomic neighborhoods relative to the nationwide representative sample.

Intervention: Think!Think!

In our intervention, we used the software called "Think!Think!" originally developed by Wonder Lab, taking full advantage of their 25 years of experience running cramming schools for school-aged children in Japan. Think!Think! has been used by approximately 1.2 million users across 150 countries and contains 15,000 individual test questions. The software is specifically designed to develop the foundational math competencies for preschool and younger primary school students, especially in the areas of spatial comprehension, logical thinking, and the concept of numbers (see, Figure 1 for an example of a Think!Think! problem).

The novel features of Think!Think! must be highlighted. It incorporates adaptive learning by leveraging an original algorithm and provides high quality problems, timely feedback, and instructions, all reflecting the students' level of proficiency. Think!Think! can achieve a TaRL approach even in a classroom with mixed ability students. Furthermore, it is designed to not only deepen the understanding of math but also stimulate children's motivation and self-esteem (e.g., through making them feel "I can do it!" and "I can try more difficult problems!"). For example, feedback

based on student performance is automatically provided, with comments that encourage students to solve more difficult problems.[3] For this experiment, Think!Think! is modified for elementary school students in Cambodia to meet local curriculum standards and was translated into the local language, Khmer.

Why did we choose to focus on math? As a number of the studies in the literature have suggested, math and science skills are highly related to economic growth across countries (Hanushek and Kimko, 2000; Jamison et al. 2007; Hanushek and Woessmann, 2016). Mathematical proficiency benefits not only to drive economic growth but also to raise individual earnings. For example, Joensen and Nielsen (2009) exploited an institutional reduction in the costs of acquiring advanced high school math in Denmark and provided evidence that the choice of a more math-intensive high school specialization has a positive causal effect on future labor market earnings. Furthermore, as Murnane and Steele (2007) showed, math teachers have been scarce because schools must compete with the better-paid private sector.

Class-Level Clustered Randomized Controlled Trial

If we allowed students to access CAL based on their own preferences, the software would most likely be used by higher-achieving students. Random assignment of access to the CAL-based software avoids this selection bias. However, because the random assignment of individuals to a particular treatment condition can be very difficult to implement in public schools. It is more common to use clustered RCTs in education setting. If the program is scaled up more broadly in future, it may make sense to implement it at the classroom, school, or district level, rather than at the individual level.

Therefore, we randomized students within intact classrooms, rather than individual students within the classrooms. Because target schools have several classes in each grade, we randomly picked two classes from each school.[4] We then employed a stratified randomization by class to ensure a

[3] One of the local governments in Japan, Mie Prefecture, has introduced Think!Think! into its public primary education system. In 2018, 17 schools with 869 students used Think!Think!. According to the survey conducted with those students, 64.6% of students responded that they wanted to attempt solving more difficult problems before they started using Think!Think! in June. By December, after six months of the program, this number had increased to 93.1%, which suggests that exposure to Think!Think! may improve students' motivation to study, although it cannot be said that the effect is causal because of the absence of a counterfactual counterpart.

[4] Wat Krous, Phum Thom, and NGS-Anuk Wat each had only two classes in each grade, whereas

balanced sample: i.e., we randomly picked one class as a treatment group from each grade (grades one to four) at each of the five schools, enabling us to create 20 treatment and 20 control classes.

One issue is that our experiment may suffer from small numbers of clusters. Interviews with the school principles of the target schools indicated that students were randomly assigned into classes within schools at the beginning of each academic semester. Thus, our randomization can be considered as being close to the student-level randomization. The classroom-level intracluster correlation coefficients for IQ score at the baseline survey is 0.379. If we assume a statistical significance of less than 0.05 for a two-tail test, the effective sample size is approximately 32 clusters with a statistical power of 80. Furthermore, our statistical inference is based on a wild-cluster bootstrap procedure developed by Cameron and Miller (2015) to provide a precise estimation of p-values, even with a small number of clusters. Thus, despite of the small numbers of clusters, we consider that our estimation successfully detects meaningful effects.

Our experiment had two phases: the first phase ran for three months from May to August 2018 and the second phase ran for seven months from January to July 2019. This setting enables us to evaluate the impact on student outcomes in both the short and longer run. Table 1 summarizes the timeline of our experiment.

First Phase (May to August 2018)

The first phase of the intervention took place over three months from May to August 2018. We had a pool of 1,654 students in five schools who were assigned either into a treatment group (834 students) or control group (820 students). Students in treatment classes were allowed to use Think!Think! for approximately 30 minutes, six days a week during their math class and were given free access to a tablet individually when using Think!Think!. Students in the control classes took the math classes led by teachers as usual.

Second Phase (January to July 2019)

The experiment continued into a second phase, which took place over

---

Prek Russey and Kroper Ha had four to six classes per grade.

seven months from January to July 2019[5]. At the beginning of the new semester, the students were randomly reassigned within schools into classes. However, because 99 students repeated a grade, 77 dropped out, and 16 transferred to other schools, only 1,460 students out of our initial 1,654 moved up to the next grade. Moreover, because several target schools had more than two classes per grade, 293 students were not reassigned into either a treatment or a control class in the new academic year, a process which occurred at random. As a result, the total numbers of participants in the second-phase intervention was 1,167.

This setting created three groups for us to compare: (i) 340 students who were assigned into treatment classes for the entire 10 months of the experiments (i.e., in both phases); (ii) 264 students who were first assigned into the treatment group during the first phase of the intervention and then into the control group in the second phase; and (iii) 563 students who were assigned into the control group for the full 10 months. In creating these three groups for comparison, our objective was not only to estimate the longer-run effect of the program but also to investigate whether the initial gains persisted after students left the program.

Baseline and Follow-up Surveys

Prior to the first-phase intervention, we conducted baseline surveys in class from May 21 to 28, 2018, with the full surveillance of teachers and staff. The baseline survey included a 30-minute paper-and-pencil IQ test and a 20-minute survey. Both were conducted in the morning, the IQ test first and then the survey. The test was translated into the local language, Khmer, and was modified appropriately for the local environment (e.g., illustrations portrayed local banknotes, food, and people of Cambodian appearance).

Following the three-month intervention in the first phase, a follow-up survey was conducted from August 16 to 25, 2018. Again, we administered the 30-minute IQ test, and the survey of students, focusing only on time-varying variables. After the second-phase intervention, another follow-up survey was conducted from August 2 to 12, 2019.

II. Data, Balance Test and Summary Statistics

[5] Schools were closed during the period from April 3 to 21 for the Cambodian New Year's Day holidays.

Variables Coded

The outcome variable of interest is defined as follows. To measure students' cognitive skills, we use the new Tanaka B-type Intelligence Scale (Tanaka et al., 2003), which has long been used in Asian countries as an age-appropriate measure of young children's IQ, and has demonstrated a high level of reliability[6]. This scale is converted to mental age and then IQ scores are calculated as mental age divided by chronological age multiplied by 100. This scale is characterized by the fact that it does not depend heavily on language skills, but instead attempts to measure IQ by having children solve problems that are most likely to be correlated with the standardized achievement tests in math, such as spatial comprehension and patten recognition. Our data suggests that this IQ scale is correlated with the standardized math scores (alpha = 0.39).[7] The Cronbach alpha for this measure is 0.61 for the baseline survey.

The survey also included a set of questionnaires to measure students' noncognitive skills. We use two psychological scales. The first scale, originally developed by Sakurai and Takano (1985), measures student's motivation in classroom. The total score is calculated by using 30 questionnaires and consists of six subscales relating to: (i) curiosity/interest, (ii) internal perceived locus of causality, (iii) independent mastery attempts, (iv) objective, (v) preference for challenge, and (vi) enjoyment (Appendix II provides more detail). The second scale, developed by Rosenberg (1985), measures students' self-esteem. Both scales are widely used in the fields of both economics and psychology.[8] For the noncognitive

[6] It is more common to use the Wechsler Intelligence Scale for Children (WISC) to test the overall intelligence. However, WISC requires special equipment, a trained proctor, and longer hours for assessment. The new Tanaka B-type Intelligence Scale is more practical as a test of intelligence for groups of students in classroom settings. It has been confirmed that there is strong correlation between the WISC and the new Tanaka B-type Intelligence Scale (Uno et al., 2014).

[7] We collected data on standardized math scores only for grade three and grade four students at the first phase of intervention. Grade three students took the National Assessment Test of Cambodia and grade 4 students took the Trends in International Mathematics and Science Study originally administered by the International Association for the Evaluation of Educational Achievement. We allowed access to the past exams, which the students in our intervention had never taken previously. However, because the grade-appropriate standardized tests were only available for grade three and grade four, we used the new Tanaka B-type Intelligence Scale as the primary outcome to measure students' cognitive skills (Ito et al., 2020).

[8] Recent research has shed light on the role of noncognitive skills in accumulating human capital (Heckman and Rubinstein, 2001). For example, Bowles et al. (2001) surveyed 25 articles written from

measure, we construct a single measure by summing all the relevant items after correcting reverse-coded items. The Cronbach alphas for these two scales for the baseline survey are 0.65 and 0.49, respectively, indicating that the set of items are modestly related as a group. All cognitive and noncognitive outcome measures are standardized to have a mean of zero and a standard deviation of one in the baseline, when we run the regression analysis.

The survey asked students to provide demographic information, such as gender, grade, month of birth, and hours of study at home. We also used the hours of studying at home (expressed in minute) as an outcome variable. 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 hours and the maximum equal to four hours, and then took the median value for the categories between two (= less than 30 minutes) and five (= 2–3 hours).

In addition, we asked the parents of students to respond to a survey, which included a set of questions about their socioeconomic background. The response rate for the parental survey was 85%. The variable on parental education represents the highest level of education of either one of the parents.

Balance Check

Table 2 shows a balance check performed for this study. The mean of the Tanaka B-type IQ test score is not statistically different at the 5% significance level between treatment and control students, and this also applied to self-esteem and motivation. The demographic variables, such as gender, age, and parental educational backgrounds, are not significantly different between two groups. Thus, it can be said that treatment and control groups that we created are successfully balanced.

III. Results

---

the late 1950s to the early 1990s that study the effect of noncognitive skills on wages. They showed that a substantial portion of the return to schooling was generated by noncognitive capacities. Cunha et al. (2010) found that 16% of the variation in educational attainments is explained by cognitive capabilities, 12% by noncognitive capabilities, and 15% by parental investment. Jokela et al. (2017) suggested that noncognitive skills have been more valuable for the past 15 years and that they predict subsequent income later in life.

Econometric Specification

To identify the causal effect of using CAL in class, we conduct an analysis of covariance (ANCOVA) using the following model and identify the effect of using CAL[9]. Our equation of interest is:

$$Y_{ijt} = \alpha + \beta T_{ijt-1} + \gamma Y_{ijt-1} + \epsilon_{ijt} \qquad (1)$$

where Y_{ijt} is the outcome variable of student i in school j at time t. The key independent variable of interest is T_{ijt}, which is a dummy variable coded as one if a student is assigned to a treatment class, and zero otherwise. e_{ijt} is the idiosyncratic error term. The crucial identifying assumption in this empirical model is that the relationship between exposure to the CAL-based software and students' unobserved ability is 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 CAL-based software on student outcomes.

Effect on Cognitive Skills in the Short run

To begin, we examine the short-run effect of CAL by using the data from the first phase of the intervention. The tables report the ordinary least squares estimates, along with block bootstraps of standard errors.[10] Model 1 provides ANCOVA estimates after controlling for the dummy variable coded as zero if the data in the baseline survey is missing and stratum fixed effects.[11] Model 2 further controls for the basic demographic controls, including gender and parental education. As adding a set of control variables to Model 2 does not changes the magnitude of the coefficients nor improve the precision of our estimates in explaining the variation in outcome variables, hereafter we interpret our results based on our preferred point

[9] According to McKenzie (2012), ANCOVA is preferred for experimental designs over a difference-in-difference approach when the autocorrelation in outcome variables between the baseline and the follow-up survey is low. Because our results exhibit weak autocorrelation in outcome variables between baseline and follow-up surveys, as demonstrated in Tables 3, 4 and 5, we apply ANCOVA for our estimation.

[10] The clustered standard errors can be used to correct for any unobserved correlations between the outcome of students in the same classroom. However, because there are only 40 clusters in our experiments, the inference may be cluster-biased (Angrist and Pischke, 2008). We address this concern by using wild bootstrapped clustered standard errors, as suggested by Cameron and Miller (2015).

[11] In total, 7.3% of target students participated in the follow-up survey only.

estimates derived from Model 1.

The CAL program appears to be successful. Table 3 ("short-run effect") shows that the estimated coefficient on IQ in the short run are positive and statistically significant at the 1% level. This result indicates that exogenous exposure to the CAL for three months raises students' average IQ scores by 0.552 SD. The magnitude of the increase in cognitive skills appears to be very large, compared with evidence presented in many of the existing studies. In particular, this is by far one of the largest effect sizes reported not only among in-class CAI programs but also out-of-class programs. Our results provide an affirmative answer to the question of whether in-class CAI program is worthwhile substituting traditional classroom instructions, addition to the literature showing that out-of-class program is effective.

We note that our point estimate in the first phase of intervention is equivalent to the result reported in Muralidharan et al. (2019), who applied a similar educational software called "Mindspark" to an intervention involving relatively poor students in Delhi, India. The comparable instrumental variable estimates in their paper indicated that lottery-winner treated students scored 0.60 SD higher compared with control students for 90 days, although their outcome variables were measured by standardized test scores.[12]

The fact that our program and that of Muralidharan et al. (2019) can boost IQ scores significantly may be helpful in understanding why some CAL programs are more effective than others. We find that Think!Think! and Mindspark are very similar. Muralidharan et al. (2019) highlighted that the advantage of using Mindspark are (i) high quality instructional materials, (ii) the adaptive content, (iii) the ability to alleviate a student-specific conceptual bottleneck, and (iv) the interactive user interface. These are also features that characterize Think!Think!. Such features may mitigate the problem of teaching in class where student achievements and abilities are largely heterogeneous by enabling TaRL. As described later, we find that Think!Think! is equally effective for all levels of initial proficiency.

There are also several differences in experimental setting between our study and Muralidharan et al. (2019). Most importantly, their intervention was conducted as an *out-of*-class remedial education. Moreover,

[12] We ran the same specification with Model 1 in equation (1) and obtained coefficients on standardized math scores: 0.767 for grade three and 0.681 for grade four, respectively (Ito et al. 2020).

it was a "blended learning" program, meaning it involved "a combination of the Mindspark computer-aided learning program, group-based instruction and extra instructional time" (p. 1429). Because Think!Think! does not require any additional lectures and instructions from teachers, our study is less likely to suffer from confounding factors, such as teacher quality.[13]

Effect on Cognitive Skills in the Longer run

Now, we proceed to examine the longer-run effect by using the data from the second phase of the intervention. The empirical specification to estimate the longer-run (β_1) and persistent (β_2) effects is:

$$Y_{ijt} = \alpha + \beta_1 LT_{ijt-1} + \beta_2 PT_{ijt-1} + \gamma Y_{ijt-1} + \epsilon_{ijt} \qquad (2)$$

The longer-run effect in Table 3 shows that the estimated coefficient on IQ in the longer run is positive and statistically significant at the 0.1% level. It suggests that students undertaking the seven-month treatment involved their IQ scores by 0.699 SD. Based on Welch's t test, the effect size in the longer run is statistically indistinguishable from the one in the short run (t = 0.02). Thus, it can be said that the additional seven-month treatment maintains the academic gains that we found for the short-run intervention.

Our results are consistent with Banerjee et al.'s (2007) school-level clustered RCT in India, which showed that CAL raised standardized math scores by 0.35 SD in the first year and 0.47 in the second year, although we note that our point estimates are larger than those of Banerjee et al. (2007). This finding brings us to another research question of interest, namely whether an initial gain persists over time and last beyond the period of the intervention. To determine the answer to this question, we must use the 2019 follow-up survey to compare the impact of the intervention on the three-month treated students with the control students. The results in Table 3 ("persistent effect") indicate that the estimated coefficient drops sharply by 0.085 SD and is statistically insignificant.[14] This result suggests that the

[13] Mindspark provides Hindi (language) programs as well as math for middle school students (grades six--nine), whereas Think!Think! specializes in math for younger primary school students (grades one--four).

[14] Our results are robust when the standardized math score measured by the National Assessment Tests of Cambodia is used as outcome variable instead of the IQ score. Thus, the quick decline in our program effect is not caused by a difference in content between Think!Think! and the national standard math curriculum in Cambodia.

program effect does not last beyond the three months for which the students were undertaking the program. The results shown in Banerjee et al. (2007) are consistent with this. They showed that one year after the CAL program, the effect on math achievement dropped significantly, by about one third, or 0.09 SD.

Except for Banerjee et al. (2007), there are few studies that investigate whether the effect of educational programs is persistent. One exception is Abeberese et al. (2014), who experimentally evaluated a one-month read-a-thon program and found that the program improved students' reading skills by 0.13 SD. Similar to our results and Banerjee et al.'s (2007), the effect fell to 0.06 SD three months after the program ended. Combined these findings indicate that the initial academic gains obtained as a result of an educational intervention targeting cognitive skills are not lasting after students leave the program. Continuous investments into children's human capital may be the key to persistent and cumulative academic gains.

Effect on Noncognitive Skills

Another outcome of interest is noncognitive skills. In contrast to cognitive skills, we do not find any evidence that our intervention improves children's noncognitive skills in the short run, as shown in Table 4 and 5. However, the effects become significant when students continue to be exposed to CAL: the coefficients on motivation and self-esteem become 0.371 and 0.265 SD, respectively. Among the six sub-scales of motivation, "objective" is large and statistically significant (see Appendix II). This subscale indicates that students' motivation in the classroom has altered; rather than being driven by extrinsic motivations (i.e., praise from parents and teachers), they are inspired by intrinsic motivations[15] (i.e., learning because they want to know more).

It is surprising that the effect on noncognitive skills appears to persist after students have left the program. One possible explanation of why the positive effects appear to become significant in the longer run is that improvement in students' short-run performance has a lagged effect on

[15] Based on self-determination theory (Deci and Ryan, 1985), intrinsic motivation refers to being engaged in a certain activity based on one's interest rather than someone's contingencies (Ryan and Deci, 2000).

noncognitive skills.[16] Because students did not receive any feedback on the IQ and achievement test scores measured by the surveys, they would have self-assessed their own academic competence, perhaps based on their classroom comprehension and/or the reactions from their teachers. Therefore, there may be a lag in the effect on the self-reported noncognitive measures compared with the cognitive ones. Consistent with this, Gonida et al. (2006) provided empirical evidence using a lagged regression analysis based on longitudinal data that actual student achievements were a strong predictor of the subsequent perceived academic competence. Many of psychological studies have demonstrated that perceived competence affects intrinsic motivation and self-esteem (Wigfield et al. 1994; Tafarodi and Swann, 1995; Bouffard et al. 2003). Thus, our findings suggest that an improvement in actual academic performance has led to improvements in perceived academic competence, which in turn has led to improved intrinsic motivation and self-esteem.

To prove this point more clearly, we add the IQ score measured at the first follow-up survey conducted in 2018 into the regressions. This allows us to evaluate how an exogenous increase in IQ scores among treated students during the first-phase intervention affects the subsequent noncognitive skill formation. Although the coefficients on motivation and self-esteem do not change even after controlling for the baseline IQ score, Table 5 shows that the coefficients are reduced substantially after controlling for the follow-up IQ score. The estimates imply that about 30% of the longer-run difference in noncognitive skills is explained by IQ scores.

In sum, the effect of the intervention on students' noncognitive skills remains significant, even though the effect on cognitive skills diminishes after the program ends. The recent studies have suggested that noncognitive skills fostered in classroom environment is persistent in the longer run (Alan

[16] The direction of causality between motivation and student achievement remains very controversial. Some studies have shown that motivation predicted subsequent student achievements (Areepattamannil et al., 2011), whereas others did not find such evidence (Marsh et al., 2005). Some studies have found that motivation and math achievement are mutually related over time (Corpus et al., 2009). The causality between self-esteem and student achievement has not been yet established. Based on a review of prior studies, Marsh and Craven (2006) found that student performance and self-esteem are both a cause and an effect of each other. However, in an influential review, Baumeister et al. (2003) concluded that self-esteem has little or no causal impact on subsequent academic performance. Rather, they suggested that the relationship may be reversed, i.e., the better performance would lead to higher self-esteem.

and Ertac, 2018; Alan et al., 2019). The randomized experiments involved training programs that helps teachers to foster students' noncognitive skills, such as self-control, patience, and grit, through structured curriculum and showed that treated students significantly improved these skills. The effects extend to actual achievement outcomes and persist almost 2--3 years after the intervention.

Unfortunately, we cannot provide any concrete evidence on how long this noncognitive gains persists, unless we are able to track students down in a few years' time, which may be beyond the scope of this study. However, we believe that our results are encouraging because we find that CAL is effective because it not only increases children's cognitive skills but also enhances their noncognitive skills, and the effect remains in almost a year after the end of the program.

Heterogeneity

When we include the interaction term and test for heterogeneous effects for gender and parental education, we obtain small point estimates on nearly all the interaction terms (see, Appendix I). IQ and motivation seem to be slightly higher for girls in the longer run, but other interaction terms are statistically insignificant.[17] Contrary to the previous literature, our results do not support the hypothesis of significant heterogeneous effects. Note that the achievement gains are homogeneous and benefited at all levels of the achievement distribution equally. This may be good news for policy-makers because it is not necessary to develop CAL program tailored for a particular subgroup of students in the public education system.

Threat to the Internal Validity and Robustness Check

Hawthorn and John Henry Effects

A drawback of our study is that evaluation-driven behavioral changes

[17] Sawada et al. (2019) empirically estimated the effect of the *Kumon* method, involving self-learning at the right level, on grade three and grade four students' cognitive and noncognitive skills in Bangladesh. Although their intervention was not computer-based, the Kumon method is aligned with the TaRL philosophy and approach. Sawada et al. 's (2019) intervention did not improve average noncognitive skills measured using the self-esteem scale. However, there were heterogenous effects, and these were largest for students with high initial level of cognitive and noncognitive skills. They also found a catching-up effect: i.e., there was a positive effect on the self-esteem scale for student with low initial cognitive and noncognitive skills.

may exist in the treatment group (the Hawthorn effect) and/or in the control group (the John Henry effect). The John Henry effect boosts the outcomes among students in the control group. If this effect exists our estimates of positive effects among the control group may be underestimated. However, such an effect would have little impact on the conclusion of this paper.

Conversely, the Hawthorn effect artificially improves students' outcomes in the treatment group. In this case, our results may be overestimated, which is more serious concern. To check whether the Hawthorn effect is present, we examine the number of blank response rates in the surveys, as a proxy for the extent to which respondents exerted their best effort in undertaking the surveys. Fortunately, there are no significant differences in the blank response rates to the follow-up survey between the treatment and control groups (we find a mean of 25.9% with 10.3 SD in the ten-month, longer-run treatment group; 26.9% with 9.10 SD in the first three months only, short-run treatment group; and 26.5% with 9.59 SD in the control group). Thus, we consider that the Hawthorn effect is minimal and unlikely to be a concern in our experiment.

Data Attrition

Because our intervention was implemented during the regular math classes and covered students in five public primary schools, our data is less likely to suffer from attrition or noncompliance biases than is the case for prior studies. However, in our study, 77.2% of target students in the first-phase intervention and 81.3 % in the second-phase intervention participated in both baseline and follow-up surveys, and the sample attrition is a significant threat to the comparability of treatment and control groups.[18]

If our intervention is successful, the low-achieving students assigned to the treatment group may not drop out during the intervention, whereas their counterpart low-achieving students assigned to the control group may drop out of school altogether. In this case, the estimated impact of this intervention may be downward biased. We calculate the attrition rate for both

[18] In the first-phase, 6.3% of students participated in the baseline survey only, 7.3% participated in in the follow-up survey only, and 9.2% participated in neither survey. We used a dummy variable coded as zero if the data in the baseline survey are missing for the estimation. We know very little about 9.2% of students who completed neither the baseline nor the follow-up survey. According to the administrative data provided by target schools, 4.6% of students were dropped out from schools and 0.9% transferred to other schools during the intervention in 2017-18.

treatment and control groups and checked whether the different characteristics of students who dropped out of the two groups differed. Fortunately, there is no evidence of differential attrition rates and different types of attrition in the treatment and control groups.

Additionally, we test the robustness of our results to attrition by modeling the selection into the follow-up based on observed characteristics, such as gender, month of birth, parental education, and grade-by-school fixed effects, and show the inverse probability weighted treatment effect (Table 7).

Peer Effects

Peer effects may be a potential threat to the internal validity of this experiment, i.e., interactions between students in the treatment and the control classes may occur and result in spillover benefits. To reduce the risk of such spillovers, we allowed the treatment students neither to access CAL outside of class and or to take their tablet home.

However, our class-level clustered randomization may not be enough to contain the spillovers between treatment and control classes. The unbiased estimate may be larger if there is a positive spillover within treated and control students, which would have little impact on the conclusion of this paper. We also cannot rule out the possibility of negative spillovers. For example, control students may concern that they are exposed by older fashioned instruction, which negatively affect their academic performance in class. If it is the case, our estimate may be overestimated, which is more serious concern. To investigate this issue, we test the effect of spillovers by comparing the mean IQ score between control students and students from Prek Pra elementary school where students took the surveys although the school is not eligible for our intervention. Figure 2 clearly shows there is no significant difference. Thus, we consider that the spillovers are also minimal and unlikely to be a concern in our experiment.

Quality of Classroom Teachers

Another threat to internal validity is violation of the stable unit treatment value assumption owing to the heterogenous quality of classroom teachers. To avoid this situation, we hired full-time instructors from local communities and placed them in the classrooms rather than the usual classroom teachers during the time that students accessed Think!Think!.

Because Think!Think! does not require additional lectures or instructions, the role of our instructors was limited to assisting students with technical matters and time management. Thus, we can rule out that the out results are confounded by teacher and/or instructor quality.

Private Tutoring

In Cambodia, students attend either morning classes (7:00--11:00 am) or afternoon classes (1:00--5:00 pm). Because some of students attend informal education programs before or after class, this may be the potential confound the results. Consequently, we obtained information on whether students attended the extra tutoring in any subject during the intervention period. Our results are robust even after controlling for attendance at informal education programs. The results are available upon request.

IV. Discussion, Mechanism, and Policy Implications

In this section, we discuss the mechanism explaining why the impact of CAL on cognitive and noncognitive skills is substantial. While several researches have suggested that instructional time is vital in the educational input of education production function (Marcotte, 2007; Gershenson and Tekin, 2018)[19], our experimental setting successfully disentangles the effect of CAL itself from longer instructional hours at school.

Although we can rule out the possibility that our results are confounded by extra instructional time, one possible concern is that our intervention may alter students' studying habits at home. To investigate this issue, we can use the information from the parental survey to estimate the effect of CAL on time spent studying at home. The coefficients are robust to be statistically insignificant at the 0.1 percent level both for short- and longer-run outcomes across all specifications (Table 8). This result is not surprising. As treatment students used tablets only during math classes, we did not expect spillover effects on studying habits at home. In other words,

[19] Marcotte (2007) used the variation in school closures due to severe winter weather in Maryland to investigate the impact of instructional time on students' cognitive skills. Gershenson and Tekin (2018) identified the causal impact of a reduction in instructional time using the geographical variation in school proximity to the "Beltway Sniper" attack in Virginia in 2002. The result also demonstrated that school closures and student absence resulting from this tragic event reduced students' math achievements by around 2--5% a year.

we do not find any significant effect in terms of treatment students studying longer hours either at school or at home. Taken together, our results suggest that CAL improves student's *productivity per hour spent studying*, rather than the results arising from increased hours for instruction at school or studying at home.

In addition, we use the attendance records during interventions and present the returns on an extra day of attendance. We define the attendance variable as the number of days that a treated student logged into Think!Think! during the intervention, with control students corresponding to zero attendance, and we then regress the attendance on the outcome variables. Because attendance may be endogenously determined, we instrument for attendance with the randomized CAL assignments. The result in Table 9 shows that, on average, the additional days of being exposed to CAL increased IQ by 0.009 and 0.006 SD in the short and longer run, respectively. Based on the Welch's t test, the effect size in the longer run is statistically indistinguishable from that in in the short run ($t = 1.00$). Attending one school year, which is roughly 180 instructional days, would lead to very large academic gains of 1.08--1.62 SD.

It is evident that the effect of our program is very significant compared with other experimental studies on CAL to date. There are several reasons why our program exhibits such a large effect. Firstly, Think!Think! better is designed to guide students' personalized learning and achieve a TaRL approach. Secondly, our program is very intensive. As mentioned earlier, treatment students undertook the program for 30 minutes per day on six days of the week during the intervention period. Third, the quality of teaching in classroom as counterfactual may be very low. A majority of teachers in our target schools are less experienced and motivated. During the intervention, we have often seen it happen that young teachers bring their own infants to classroom and cuddle them while teaching. The large gains from our program may be attributed to replace the low quality of teaching with highly personalized computer assisted instruction.

As well as finding a large effect on average, our results do not provide any evidence of significant heterogeneity of effects in terms of students' demographic characteristics and initial achievement levels. Thus, policy-makers do not need to be concerned that CAL exacerbates the existing educational inequalities. Moreover, Think!Think does not impose any

additional workload on teachers. Taking all these advantages into account, we consider that scaling up this program and implement it in other locations is realistic and promising proposition.

V. Conclusion

In this paper, we examine the causal effect of an in-class CAL program on children's cognitive and noncognitive skills. We ran a clustered RCT at five elementary schools near Phnom Penh over two years, involving 1,657 students from grade one to four. Students were randomly assigned to either one of 20 treatment classes or one of 20 control classes. Treated students were allowed to access CAL during their regular math classes for three months.

When new academic year commenced, students were reassigned either into treatment or control classes for the additional seven-month comparison. As a result, we had three groups of students to compare: (i) students who were assigned to treatment classes for the entire 10 months; (ii) students who were assigned to treatment classes for the first three months and then control groups for the remaining of seven months; (iii) students who were assigned to control groups for the entire 10 months.

Most importantly, we found that the average treatment effects on cognitive skills are positive and statistically significant. Our preferred point estimates of the impacts on IQ are 0.552 SD in the short run and 0.699 SD in the longer run. Improvements in students' cognitive skills are most likely attributable to the increased learning productivity per hour, not to the increased hours of instruction at school or study at home. Second, the initial academic gain sharply drops by 0.085 SD and becomes statistically insignificant. Combined with findings from the previous literature, it appears that the initial gains in cognitive skills obtained as a result of an educational intervention do not persist after students leave the program. Conversely, the meaningful effect on noncognitive skills is not detected immediately after the first three-month short-run program, but it appears to become significant in the longer run and persists even after students leave the program. Because the estimated coefficients decline substantially after controlling for IQ scores, which partly contribute to improve subsequent improvements in noncognitive skills.

In contrast to the previous literature, our results do not support the hypothesis of significant heterogeneous effects of CAL. This may be good news for policy-makers because it suggest that developing CAL programs tailored to a particular subgroup of students within the public education system unnecessary. CAL may be a promising option to scale up the TaRL approach in other locations.

Figure 1: Example of a problem in Think!Think!

(Source) Wonder Lab

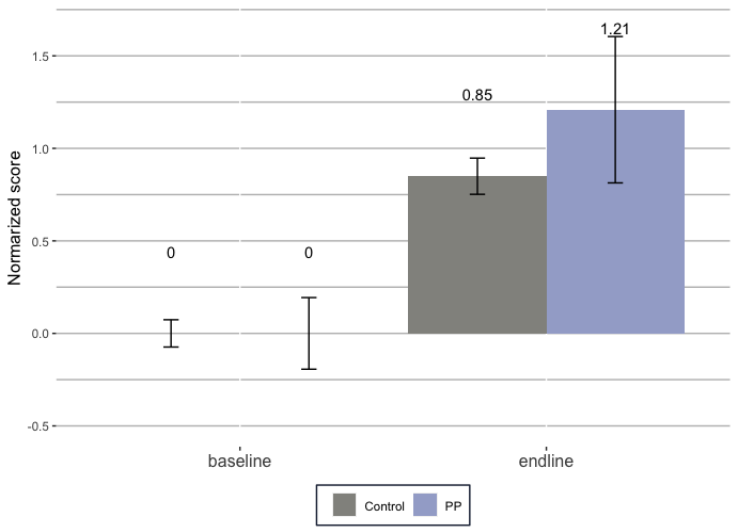

Figure 2: The Spillovers in IQ

(Note) This figure compares the mean IQ score of control students with students from Prek Pra elementary school where students took the surveys although their school is not eligible for our intervention. The IQ scores are standardized to have a mean of zero and a standard deviation of one in the baseline (in August 2018 for control students and January 2019 for Prek Pra students).

Table 1: Timeline

Baseline Survey	May 2–28, 2018
First-Phase Intervention	May to August 2018 (three months)
Follow-up Survey	August 16–25 2018
Reshuffle Treatment/Control Classes	January 2019
Second-Phase Intervention	January to July 2019 (seven months)
Follow-up Survey	August 2–12, 2019

Table 2: Balance Check and Summery Statistics

	Treatment (**T**)	# of obs	Control (**C**)	# of obs	Difference (**T**)-(**C**)
Tanaka B IQ	78.795	685	78.404	706	0.375
	(13.777)		(13.137)		(0.623)
Motivation	21.024	338	21.004	272	0.042
	(3.630)		(3.641)		(0.273)
Self-esteem	2.818	334	2.870	271	-0.051
	(0.377)		(0.415)		(0.032)
Studying at home	167.481	526	169.882	424	-2.789
(in minutes)	(123.173)		(108.978)		(7.902)
Gender	0.481	829	0.473	816	0.008
(male = 1)	(0.500)		(0.500)		(0.025)
Age	8.201	816	8.203	807	-0.016
	(1.528)		(1.630)		(0.052)
Parental education					
College or above	0.029	631	0.018	619	0.011
	(0.167)		(0.132)		(0.008)
Higher secondary	0.445		0.480		-0.037
	(0.497)		(0.500)		(0.027)
Lower secondary	0.304		0.286		0.024
	(0.460)		(0.452)		(0.026)
Primary	0.222		0.216		0.002
	(0.416)		(0.412)		(0.023)

(Note) Treatment and control refer to whether students are randomly assigned into classes with CAL and classes without CAL, respectively. The variables used in this table are from the baseline survey conducted in May 2018. The numbers reported in each cell represent means along with standard deviations. The column "Difference" shows the estimates drawn from regressing the outcomes on a treatment dummy coded one if students are randomly assigned into classes with CAL and stratum fixed effects.

Table 3: Effect on IQ Score

	First phase of intervention		Second phase of intervention	
	Model 1	Model 2	Model 1	Model 2
Short-run effect (for three months)	0.552*** (0.066)	0.543*** (0.066)		
Longer-run effect [β_1] (for entire 10 months)			0.699*** (0.088)	0.671*** (0.072)
Persistent effect [β_2] (for first three months only)			0.085 (0.121)	0.089 (0.108)
Baseline	0.497*** (0.023)	0.509*** (0.023)	0.443*** (0.046)	0.439*** (0.044)
Stratum fixed effects	✓	✓	✓	✓
Controls		✓		✓
# of obs.	1,410	1,158	1,029	841
Adj-R^2	0.435	0.437	0.320	0.324

(Note) Treatment is a dummy variable coded as one if a student is assigned into treatment classes. Model 1 controlled for prior score, missing baseline dummy and stratum fixed effects. Model 2 further controlled for gender, month of birth and parental education using the information from the parental survey, for which the response rate was 85%. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors, as suggested by Cameron and Miller (2015), are given in parentheses.

Table 4: Effect on Motivation

	First phase of intervention		Second phase of intervention	
	Model 1	Model 2	Model 1	Model 2
Short-run effect (for three months)	0.009 (0.076)	-0.038 (0.080)		
Longer-run effect [β_1] (for entire 10 months)			0.371*** (0.109)	0.252*** (0.132)
Persistent effect [β_2] (for first three months only)			0.283** (0.121)	0.373*** (0.116)
Baseline	0.510*** (0.027)	0.491*** (0.034)	0.336*** (0.039)	0.291*** (0.037)
Stratum fixed effects	✓	✓	✓	✓
Controls		✓		✓
# of obs.	991	826	973	797
Adj-R^2	0.160	0.201	0.158	0.152

Table 5: Effect on Self-Esteem

	First phase of intervention		Second phase of intervention	
	Model 1	Model 2	Model 1	Model 2
Short-run effect (for three months)	0.087 (0.076)	0.074 (0.057)		
Longer-run effect [β_1] (for entire 10 months)			0.265*** (0.084)	0.208* (0.109)
Persistent effect [β_2] (for first three months only)			0.145 (0.107)	0.269*** (0.071)
Baseline	0.085** (0.040)	0.076** (0.032)	0.071 (0.046)	0.093** (0.045)
Stratum fixed effects	✓	✓	✓	✓
Controls		✓		✓
# of obs.	1,400	1,132	1,029	841
Adj-R^2	0.091	0.089	0.062	0.050

(Note) Treatment is a dummy variable coded as one if a student is assigned into treatment classes. Model 1 controlled for prior score, missing baseline dummy and stratum fixed effects. Model 2 further controlled for gender, month of birth and parental education using the information from the parental survey, for which the response rate was 85%. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors, as suggested by Cameron and Miller (2015), are given in parentheses.

Table 6: The Change in Coefficients After Controlling for Baseline and Follow-up IQ Scores

	Motivation			Self-esteem		
	Table 4 Model 1	Controlled for Baseline IQ (May, 2018)	Controlled for Follow-up IQ (Aug, 2018)	Table 5 Model 1	Controlled for Baseline IQ (May, 2018)	Controlled for Follow-up IQ (Aug, 2018)
Longer-run effect [β_1] (for entire 10 months)	0.371*** (0.109)	0.372*** (0.119)	0.268* (0.123)	0.265*** (0.084)	0.302*** (0.095)	0.190** (0.079)
Persistent effect [β_2] (for first three months only)	0.283** (0.121)	0.289** (0.120)	0.255** (0.111)	0.145 (0.107)	0.103 (0.088)	0.143 (0.090)
Baseline	0.336*** (0.039)	0.342*** (0.039)	0.321*** (0.041)	0.071 (0.046)	0.072 (0.044)	0.072 (0.046)
Stratum fixed effects	✓	✓	✓	✓	✓	✓
# of Obs.	973	898	919	1,029	944	968
Adj-R^2	0.158	0.182	0.176	0.062	0.059	0.061

(Note) Treatment is a dummy variable coded as one if a student is assigned into treatment classes. Model 1 controlled for prior score, missing baseline dummy and stratum fixed effects. Model 2 further controlled for gender, month of birth and parental education using the information from the parental survey, for which the response rate was 85%. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors, as suggested by Cameron and Miller (2015), are given in parentheses.

Table 7: Inverse Probability Weighted Treatment Effects

	IQ	Motivation	Self-esteem
Short-run effect	0.542***	-0.039	-0.007
(for three months)	(0.069)	(0.086)	(0.065)
Longer-run effect	0.688***	0.286**	0.194**
(for entire 10 months)	(0.077)	(0.059)	(0.090)
Persistent effect	0.060	0.351***	0.274***
(for first three months only)	(0.093)	(0.044)	(0.065)

(Note) The results in this table are weighted by the inverse of the predicted probability of participating in the follow-up survey. The probability is predicted using a probit model with gender, month of birth, parental education, and grade-by-school fixed effects as predictors. The symbols ** and * represent significance at the 1% and 5% levels, respectively.

Table 8: Effect on Time Spent Studying at Home

	First phase of intervention		Second phase of intervention	
	Model 1	Model 2	Model 1	Model 2
Short-run effect	-0.021	-0.077		
(for three months)	(0.066)	(0.052)		
Longer-run effect [β_1]			0.013	-0.063
(for entire 10 months)			(0.098)	(0.089)
Persistent effect [β_2]			0.028	-0.053
(for first three months only)			(0.053)	(0.060)
Baseline	0.278***	0.258***	0.111***	0.108***
	(0.036)	(0.036)	(0.034)	(0.040)
Stratum fixed effects	✓	✓	✓	✓
Controls		✓		✓
# of obs.	1,303	1,069	1,017	831
Adj-R^2	0.066	0.072	0.039	0.047

(Note) Treatment is a dummy variable coded as one if a student is assigned into treatment classes. Hours spent studying at home are measured on a six-point scale (from 1 = not at all to 6 = more than 4 hours). The minimum of this variable is ranged from zero to four and then coded as the median value for categories between two (= less than 30 minutes) and five (= 2–3 hours). Model 1 controlled for prior score, missing baseline dummy and stratum fixed effects and Model 2 further controlled for gender, month of birth and parental education using the information from the parental survey, for which the response rate was 85%. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors as suggested by Cameron and Miller (2015) are given in parenthesis.

Table 9: The Returns on Attendance

		IQ	Motivation	Self-esteem
Short-run effect	Attendance	0.009***	0.004***	0.000
(for three months)	(in days)	(0.002)	(0.001)	(0.002)
	# of obs.	980	962	955
	Adj-R^2	0.443	0.172	0.112
Longer-run effect	Attendance	0.006***	0.004***	0.004***
(for entire 10 months)	(in days)	(0.001)	(0.001)	(0.001)
	# of obs.	1,029	1,028	1,029
	Adj-R^2	0.320	0.156	0.060

(Note) Attendance is defined as the number of days that a treated student logged into Think!Think! during the intervention. Each specification controlled for prior score, missing baseline dummy and stratum fixed effects. We instrument for attendance with the randomized assignment of Think!Think!. The symbols ** and * represent significance 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors as suggested by Cameron and Miller (2015) are given in parentheses.

Appendix I: Heterogenous Effect by Student Characteristics

Gender (female = 1)

	IQ	Motivation	Self-esteem
Short-run effect	0.012	-0.127	-0.136
(for three months)	(0.057)	(0.121)	(0.137)
Longer-run effect	0.380***	0.270***	0.144
(for entire 10 months)	(0.100)	(0.144)	(0.170)
Persistent effect	-0.027	0.238**	0.072
(for first three months only)	(0.073)	(0.113)	(0.145)

Parental Education (reference = completed primary school)

		IQ	Motivation	Self-esteem
Short-run effect	Lower Secondary	0.071	0.021	0.314
(for three months)		(0.126)	(0.194)	(0.263)
	Higher Secondary	0.096	0.146	-0.072
		(0.132)	(0.156)	(0.185)
	College or above	-0.544**	-0.033	-0.391
		(0.242)	(0.260)	(0.376)
Longer-run effect	Lower Secondary	-0.062	0.254	0.239
(for entire 10 months)		(0.141)	(0.212)	(0.226)
	Higher Secondary	0.139	0.251	-0.013
		(0.194)	(0.194)	(0.247)
	College or above	0.239	-0.492	-0.195
		(0.446)	(0.423)	(0.432)
Persistent effect	Lower Secondary	-0.191	0.036	0.116
(for first three months only)		(0.197)	(0.255)	(0.209)
	Higher Secondary	-0.142	0.275	0.221
		(0.213)	(0.282)	(0.185)
	College or above	0.029	-0.253	-0.245
		(0.380)	(0.341)	(0.486)

Initial Score (reference = middle tercile)

		IQ	Motivation	Self-esteem
Short-run effect	Low	0.034	-0.047	-0.146
(for three months)		(0.144)	(0.176)	(0.163)
	High	0.163	-0.108	0.164
		(0.150)	(0.215)	(0.186)
Longer-run effect	Low	-0.021	-0.196	0.016
(for entire 10 months)		(0.146)	0.163	(0.242)
	High	-0.174	-0.089	0.023
		(0.228)	0.173	(0.265)
Persistent effect	Low	0.080	-0.289***	-0.098
(for first three months only)		(0.198)	(0.100)	(0.150)
	High	-0.327	-0.535*	-0.048
		(0.211)	(0.313)	(0.246)

(Note) The tables show the coefficients on the interaction term. Each specification controlled for the covariates (either gender, parental education, or initial score), prior score, missing baseline dummy and stratum fixed effects. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors, as suggested by Cameron and Miller (2015), are given in parenthesis.

Appendix II: The Subscale of Motivation in the Classroom

	Curiosity/Interest	Internal Perceived Locus of Causality	Independent Mastery Attempts	Objective	Preference for Challenge	Enjoyment
Short-run effect	0.024	-0.016	0.005	0.022	0.035	0.094*
(for three months)	(0.041)	(0.068)	(0.050)	(0.052)	(0.059)	(0.052)
Longer-run effect	0.362***	0.160***	0.088*	0.466***	0.191***	0.145
(for entire 10 months)	(0.077)	(0.061)	(0.052)	(0.153)	(0.082)	(0.109)
Persistent effect	0.028	0.122	0.153**	0.383***	0.160*	0.196**
(for first three months only)	(0.079)	(0.088)	(0.074)	(0.134)	(0.084)	(0.042)

(Note) The motivation scale was originally developed by Sakurai and Takano (1985). The scale is calculated based on 30 items and broken down into six subscales: curiosity/interest, internal perceived locus of causality, independent mastery attempts, objective, preference for challenge and enjoyment. All questionnaires are defined by an intrinsic and an extrinsic pole, and students are asked to choose one of two response for each item. For example, students could be asked to select between (a) I only need to learn what the teacher teaches me, or (b) I'm willing to learn a lot of things. If a student chooses (a) rather than (b), he or she places more emphasis on teacher approval than on her/his own curiosity/interest. The former represents an extrinsic motivation, whereas the latter represents an intrinsic motivation in the classroom. If the student chooses the intrinsic pole, he or she will score one point, and zero otherwise. Each specification controlled for prior score, missing baseline dummy, and stratum fixed effects. The symbols ** and * represent significance at the 1% and 5% levels, respectively. The wild bootstrapped clustered standard errors as suggested by Cameron and Miller (2015) are given in parentheses.

References

Abeberese, A. B., Kumler, T. J., & Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. *Journal of Human Resources*, 49(3), 611-633.

Alan, S., & Ertac, S. (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy*, 126(5), 1865-1911.

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3), 1121-1162.

Areepattamannil, S., Freeman, J. G., & Klinger, D. A. (2011). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education*, 14(2), 233-259.

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

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

Barrera-Osorio, F., & Linden, L. 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, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy*, 1(1), 52-74.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?. *Psychological Science in the Public Interest*, 4(1), 1-44.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., & 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.

Bouffard, T., Marcoux, M. F., Vezeau, C., & Bordeleau, L. (2003). Changes in self-perceptions of competence and intrinsic motivation among elementary schoolchildren. *British Journal of Educational Psychology*, 73(2), 171-186.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 39(4), 1137-1176.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317-372.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts. NCEE 2009-4041. National Center for Education Evaluation and Regional Assistance.

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

Corpus, J. H., McClintic-Gilbert, M. S., & Hayenga, A. O. (2009). Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. *Contemporary Educational Psychology*, 34(2), 154-166.

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

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3), 883-931.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2), 109-134.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5), 1739-74.

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. Upgrading Education

with Technology: Insights from Experimental Research. *Journal of Economic Literature*.

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). Education technology: An evidence-based review (No. w23744). National Bureau of Economic Research.

Gershenson, S., & Tekin, E. (2018). The effect of community traumatic events on student achievement: Evidence from the beltway sniper attacks. *Education Finance and Policy*, 13(4), 513-544.

Gonida, E., Kiosseoglou, G., & Leondari, A. (2006). Implicit theories of intelligence, perceived academic competence, and school achievement: Testing alternative models. *American Journal of Psychology*, 223-238.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American economic review*, 90(5), 1184-1208.

Hanushek, E. A., & Woessmann, L. (2016). Knowledge capital, growth, and the East Asian miracle. *Science*, 351(6271), 344-345.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, 91(2), 145-149.

Ito, H., Kasai, K., Nishiuchi, H., & Nakamuro, M. (2020). Does Computer-aided Instruction Improve Children's Cognitive and Non-cognitive Skills? *Asian Development Review*, forthcoming.

J-PAL Evidence Review. 2019. Will Technology Transform Education for the Better? Cambridge, MA: Abdul Latif Jameel Poverty Action Lab.

Jamison, E. A., Jamison, D. T., & Hanushek, E. A. (2007). The effects of education quality on income growth and mortality decline. *Economics of Education Review*, 26(6), 771-788.

Joensen, J. S., & Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes?. *Journal of Human Resources*, 44(1), 171-198.

Jokela, M., Pekkarinen, T., Sarvimäki, M., Terviö, M., & Uusitalo, R. (2017). Secular rise in economically valuable personality traits. *Proceedings of the National Academy of Sciences*, 114(25), 6527-6532.

Lai, F., Zhang, L., Hu, X., Qu, Q., Shi, Y., Qiao, Y., Boswell, M., & Rozelle, S. (2013). Computer assisted learning as extracurricular tutor? Evidence from a randomised experiment in rural boarding schools in

Shaanxi. *Journal of Development Effectiveness*, 5(2), 208-231.
Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics*, 89(4), 721-736.
Linden, L. L. (2008). Complement or substitute?: The effect of technology on student achievement in India. Mimeo
Machin, S., McNally, S., & Silva, O. (2007). New technology in schools: Is there a payoff?. *The Economic Journal*, 117(522), 1145-1167.
Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics*, 126(2), 987-1027.
Marcotte, D. E. (2007). Schooling and test scores: A mother-natural experiment. *Economics of Education Review*, 26(5), 629-640.
Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on psychological science*, 1(2), 133-163.
Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2), 397-416.
McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2), 210-221.
Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., & Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi. *Journal of Development Effectiveness*, 6(3), 300-323.
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.
Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, 109(4), 1426-60.
Murnane, R. J., & Steele, J. L. (2007). What is the problem? The challenge

of providing effective teachers for all children. *The future of Children*, 15-43.

National Institute of Statistics, Ministry of Planning (2018). Socio-Economic Survey 2017. Retrieved May 11th 2020 from https://www.nis.gov.kh/nis/CSES/Final%20Report%20CSES%202017.pdf.

OECD (2018). Education in Cambodia: Finding from Cambodia's experience in PISA for development. Retrieved May 11th 2020 from https://www.oecd.org/pisa/pisa-for-development/PISA-D%20national%20report%20for%20Cambodia.pdf.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlled trial in Kenya. *Journal of Development Effectiveness*, 8(3), 404-422.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, 47(4), 833-878.

Rosenberg, A. (1985). METHODOLOGY, THEORY AND THE PHILOSOPHY OF SCIENCE. *Pacific Philosophical Quarterly*. 66(3-4), 377-393.

Rouse, C. E., & 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.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54-67.

Sakurai, S., & Takano, S. (1985). A new self-report scale of intrinsic versus extrinsic motivation toward learning in children. *Tsukuba Psychological Research*, 7, 43-54.

Sawada, Y., Mahmud, M., Seki, M., Le, A., & Kawarazaki, H. (2019). Fighting Against Learning Crisis in Developing Countries: A Randomized Experiment of Self-Learning at the Right Level. Available at SSRN 3471021.

Schling, M., & Winters, P. (2018). Computer-assisted instruction for child

development: Evidence from an educational programme in rural Zambia. *The Journal of Development Studies*, 54(7), 1121-1136.

Shapley, K., Sheehan, D., Maloney, C., & Caranikas-Walker, F. (2009). Evaluation of the Texas Technology Immersion Pilot: Final Outcomes for a Four-Year Study (2004-05 to 2007-08). Texas Center for Educational Research.

Tafarodi, R. W., & Swann Jr, W. B. (1995). Self-linking and self-competence as dimensions of global self-esteem: initial validation of a measure. *Journal of Personality Assessment*, 65(2), 322-342.

Tanaka, K., Okamoto, K., & Tanaka, H. (2003). *The New Tanaka B Intelligence Scale*. Kaneko Shobo.

UNESCO (2006) World Data on Education, 6th edition, Cambodia. Retrieved May 11th 2020 from http://www.ibe.unesco.org/sites/default/files/Cambodia.pdf.

Uno, Y., Mizukami, H., Ando, M., Yukihiro, R., Iwasaki, Y., & Ozaki, N. (2014). Reliability and validity of the new Tanaka B Intelligence Scale scores: a group intelligence test. *PloS one*, 9(6), e100262.

Wigfield, A., & Eccles, J. S. (1994). Children's competence beliefs, achievement values, and general self-esteem: Change across elementary and middle school. *The Journal of Early Adolescence*, 14(2), 107-138.

World Bank. (2017). *World Development Report 2018: Learning to Realize Education's Promise*. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1.