Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China

Di Mo
Stanford University
LICOS Centre for Institutions and Economic Performance, University of Leuven

Weiming Huang
Stanford University

Yaojiang Shi
Center for Experimental Economics of Education, Shaanxi Normal University

Linxiu Zhang
Center for Chinese Agricultural Policy, IGSNRR, Chinese Academy of Sciences

Matthew Boswell
Stanford University

Scott Rozelle
Stanford University

Abstract: There is a great degree of heterogeneity among the studies that investigate whether computer technologies improve education and how students benefit from them – if at all. The overall goal of this study is to assess the effectiveness of computing technologies to raise educational performance and non-cognitive outcomes and identify what program components are most effective in doing so. To achieve this aim we pool the data sets of five separate studies about computer technology programs that include observations of 16,856 students from 171 primary schools across three provinces in China. We find that overall computing technologies have positive and significant impacts on student academic achievement in both math and in Chinese. The programs are found to be more effective if they are implemented out-of-school, avoiding what appear to be substitution effects when programs are run during school. The programs also have heterogeneous effects by gender. Specifically, boys gain more than girls in Chinese. We did not find heterogeneous effects by student initial achievement levels. We also found that the programs that help students learn math—but not Chinese—have positive impacts on student self-efficacy.

Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China

The use of computer technology has become increasingly popular in education over the past decades (Barrow, Markman, and Rouse 2009; Malamud and Pop-Eleches, 2011). Studies have shown that there are many advantages of using computers in education. For example, Ebner and Holzinger (2007) found that computing technology can create intrinsically-motivating environments for students. The interaction with and immediate feedback from the computer can make the learning process a more engaging experience for students (Bakar et al. 2006) and may also increase student effort at school (Schaefer and Warren 2004). Studies in developing countries like India suggest that using computers to supplement regular teaching can compensate for the shortage of teachers or poor teaching quality (Pal et al. 2006). Computer software can provide more learning material and can be programmed to teach to different levels of students (Pawar, 2006).

Despite the popularity of using computer technology in education, there are ongoing debates about whether it can actually improve student academic achievement. In a program that uses computers to boost learning among medical students, researchers actually found a negative impact on student test scores (Vichitvejpaisal et al. 2001). Contrastingly, student math test scores improved after students used computers to study math in India (Banerjee et al. 2007). Studies suggest that different implementation strategies account for such divided outcomes (Osín 1998). For example, programs that use computers to help students with learning during regular classes (henceforth, in-school programs) or during a time that is not planned for regular teaching (henceforth, out-of-school programs) may influence student achievement differently. Research has found that in-school programs may generate negative effects on learning because they may

substitute for effective regular classes (Lai et al. 2014). In contrast, other studies have found that in-school programs complement regular teaching and create positive impacts on student achievement (Tüzün et al. 2009; Liu et al. 2006).

In addition to varied program impacts, a great deal of heterogeneity exists among studies that seek to determine who benefits more from using computer technology in education. Specifically, a consensus does not exist about the role of gender in the use of computer technology in education. There are studies suggesting that boys benefit more than girls because boys become more focused on new computer technologies. A study by (Ong and Lai 2006) argues that boys perceive more utility from computers and are more motivated to learn novel technologies than girls. However, other studies have found the opposite. Girls were shown to have gained more in cognitive achievement in classes when teachers adopted computer technology in instruction (Vogel et al. 2006). Girls also were found to have gained more in computer-supported collaborative learning (Prinsen, Volman, and Terwel 2007). The authors in the latter study suggest that the greater learning occurred because girls are more collaborative than boys and more efficient at using computers when cooperation and learning are required.

It also is not clear whether the impact of using computer technology in education varies by the initial level of academic achievement of students. On one hand, higher achievers may benefit more because they are more efficient learners of new materials (Hativa 1988; Gorjian et al. 2011). In contrast, lower achievers may improve more because they are able to use computing technologies to help them catch up (Baker, Gersten, and Lee 2002) and perhaps gain more from the feedback facilitated by computers.

Studies that examine the non-academic outcomes of educational computing programs are similarly inconsistent in their findings. For example, a positive effect on self-efficacy (which in our study we define as a person's perception of his or her ability to plan and take action to reach a particular learning goal) was observed for nursing students after they used computers to simulate how to provide better care for patients (Madorin and Iwasiw, 1999). However, another study failed to identify an impact on self-efficacy when a group of college students in the US used computer programs to learn math (Maag 2004).

Several factors may account for the variation in results we find when studying the record of computing in education. First, there is significant variation in the environments in which these studies were implemented. For example, a large number of earlier studies were implemented in developed countries such as Austria, Germany, Switzerland, New Zealand and the United States (Bakar et al. 2006; Vogel et al. 2006; Maag 2004). In recent years, more studies have been conducted in developing countries (Banerjee et al. 2007; Tüzün et al. 2009; Ong and Lai 2006). The education systems in these countries differ dramatically. Program differences may either derive from differing availability of resources, such as technical support or the quality of the computing equipment, or differing levels of teacher incentives or student motivation. In addition, the targeted populations and subjects vary substantially. The targeted populations range from primary school students (Liu et al. 2006) to professionals (Baker, Gersten, and Lee 2002). Subjects range from math (Barrow, Markman, and Rouse 2009) and language learning (Hyland 1993) to professional skills such as nursing (Maag 2004).

Second, most of the existing studies are small in scale. More than half of the studies mentioned above include fewer than one to two hundred participants. The absence of sufficient statistical power in the studies may be one of the reasons for the differing results. Few studies even try to calculate the statistical power of their analyses.

Third, studies adopt different implementation protocols. For example, programs were conducted both with or without teacher instruction (Madorin and Iwasiw 1999; Ebner and Holzinger 2007; Pal et al. 2006). The intensity of the programs has ranged from 30 minutes to one academic year (Barrow, Markman, and Rouse 2009; Gorjian et al. 2011). In many of the studies the protocols are not carefully described.

The overall goal of this study is to assess the effectiveness of using computing technologies to raise educational performance and non-cognitive outcomes and identify what program components contribute to program success. In pursuing this goal, our study seeks to answer the following questions: What impacts do programs that use computers and educational software have on student cognitive and non-cognitive outcomes? Which components matter for which type of outcomes? Which forms of implementation works the best? Are there heterogeneous effects across different sub-populations?

In this paper we seek to answer these questions by pooling the data from five randomized experiments that used computer technology to assist primary school student learning in poor areas of China. We believe the strategy of combining material from five independent studies is important since a pooled study allows us to better understand the general effects of computing technology in education as well as the heterogeneous impacts on both academic and non-academic outcomes. While the original studies are valuable in assessing the impacts of various computer-based educational programs,

previous work has shown that pooling data from several studies and stacking them together can provide more statistical power for both estimating average program impact and conducting heterogeneity analysis (Taioli and Bonassi 2002). The rise in statistical power of a pooled study is also higher than a meta-analysis that treats each study as a single observation.

By building an aggregated data set from five separate studies about educational programs using computing technology, including a total of 16856 students in 171 primary schools, we find that, overall, computing technologies have positive and significant impacts on student academic achievement in both math and Chinese. The programs are found to be more effective if they are implemented out-of-school, avoiding what appear to be substitution effects when programs are run in-school. The programs are found to have heterogeneous effects by gender. Specifically, boys gain more than girls in Chinese. In contrast, boys do not seem to differ from girls in math improvement after the program. We did not find heterogeneous effects by student initial achievement levels. Lower achievers gain as much as higher achievers from the program. We also found that the programs that help students learn math—but not Chinese—have positive impacts on student self-efficacy.

Despite the contribution of our paper, we do realize that the study has limitations. First, the programs included in the studies follow protocols in which students are instructed to only interact with the computer and their computing partner. Teachers are not part of the learning process. Indeed, by protocol teachers were not allowed to provide any additional instruction. Hence, the results of this pooled study are applicable to programs that are not designed to measure programs that encourage group interactions

among students or interactions between students and teachers. Second, one of the strengths of this study is also one that limits its external validity. All of the programs are implemented in poor schools in rural China's educational system. This suggests that our results are mainly representative of schools with poor resources in developing countries. The study may say nothing about how such programs would work in schools that are more competitive in richer, better-resourced communities (Watkins 2000).

To meet our goals and objectives, the rest of the paper is organized as follows. In the next section, we present an overview of the five individual computer assisted learning programs we analyze in this paper. In Section Three we discuss the sampling strategies, data collection processes and statistical methods of the study. In Section Four we present the analytical results that seek to answer the questions about computing technologies raised earlier in this section. Section Five concludes.

An overview of the five computer assisted learning programs

In this section we introduce the computer assisted learning programs that we have run in China between 2010 and 2012. In the rest of the paper, we call these programs our *CAL programs*. For each program we describe the specific problem addressed, the main objective of the program, the approach (in briefest terms the design of the CAL program); and the results. Importantly, in the rest of the paper we will not be redoing or reporting on the results of these analyses. Rather, we will be combining the datasets from the five projects and analyzing the data to try to answer key questions about the effectiveness of CAL in general.

The first CAL program (called the Migrant CAL Program) was targeted at narrowing the education gap that exists between students from rural areas that come with their parents to Beijing and attend private, unregulated, low-quality migrant schools (henceforth, *migrant students*) and students from urban areas that attend free and high quality urban public schools (Table 1, row 1). One of the biggest problems facing many migrant students is that they frequently fall behind (their parents often move and they are in and out of many different schools) and find it difficult to catch up. Because many migrant students fall behind in school the primary objective of the Migrant CAL Program was to provide students with remedial tutoring to help them narrow the achievement gap with regular urban children in public schools. To achieve the objective, we delivered a CAL math program to migrant students during periods of time that did not conflict with their regular math or Chinese classes (e.g., before school, during lunch, after school or during a free, study hall class). The results of the Migrant CAL experiment demonstrated that CAL significantly improved student math test score by 0.14 standard deviations.

In the second CAL program we targeted groups of vulnerable students that attend rural schools in poor mountainous regions of China. Many of these students had parents that worked in distant urban centers or lived with parents during the weekend, but, due to the remoteness of their villages, lived at school in dormitories during the week (Table 1, row 2). All of the students were ethnically Han, China's largest ethnic group (making up about 92 percent of the population). Previous work (Mo et al. 2012) shows that primary school students that live in dormitories perform less well than other students. Similar to the Migrant CAL Program, we rolled out a CAL program in these poor rural schools during after-school hours in Shaanxi Province (henceforth, Shaanxi CAL Program I) with

the goal of improving educational performance among these vulnerable boarding students. The Shaanxi CAL Program I study found that the standardized math scores of students improved by 0.12 standard deviations.

The third CAL program targeted ethnic minority students in northwest China whose academic performance is, on average, lower than that of the poor rural students in Shaanxi Province (Hannum 1999). Among the most significant barriers for the minority students is their relatively low level of Chinese language skill, as Mandarin Chinese is the medium of instruction and the language of all textbooks (Lai et al. 2012). The third CAL program was conducted in Qinghai province (henceforth, Qinghai CAL Program I), where minority families live in relatively high rates of concentration. The immediate objective of the Qinghai CAL Program I was to use CAL to help students improve their Chinese during after-school hours. This program was found to have a positive impact of 0.20 standard deviations on the standardized Chinese test scores of minority students. The program also had significant positive spillover effects on math test scores.[1]

The fourth CAL program sought to determine whether program impacts differed when CAL sessions were held during regular school hours instead of during after-school hours. One reason for examining this issue is that if a CAL program was to be scaled up across a large number of schools by the formal school system, it is possible that the program would be incorporated into regular school hours (we call this kind of program an in-school CAL program). Since in-school programs may substitute for teacher instruction

[1] In this case the spillover was a positive one. The analysis found that after treating students in CAL group with a Chinese language curriculum, math test scores also went up. The most likely causal mechanism is that in China, math textbooks are written in Chinese and math classes are taught in Chinese. Hence, it appears as if when the CAL Chinese treatment improved Chinese skills of the ethnic minority students (as we found in the analysis), math test scores also rose.

and other learning activities, it is not clear whether an in-school CAL program will help improve student learning as much as an after school program. In the Shaanxi CAL Program II, students were offered CAL sessions in both math and Chinese. Therefore, the objective the fourth CAL program (henceforth, the Shaanxi CAL Program II) was to test whether an in-school CAL program is effective in improving student test scores. The results of the Shaanxi CAL Program II showed that student math scores did improve (in this case by 0.16 standard deviations). However, no impact was found on Chinese test scores.

The fifth CAL program targeted the minority students and was designed to test whether CAL can also improve math scores directly by providing students with math in addition to Chinese sessions. The program was conducted in Qinghai Province—we call it the Qinghai CAL Program II. The objective of Qinghai CAL Program II was to test whether directly engaging minority students in math CAL sessions will help them improve even more than when they were only engaged in Chinese CAL sessions. The program was supposed to be implemented as an out-of-school program. However, during implementation, it was discovered that some of the schools implemented the Qinghai CAL Program II as an in-school program (because there was sometimes not enough out-of-school time to accommodate the program).[2] The results suggest that the Qinghai CAL Program II improved student test scores only among the schools that implemented it as an out-of-school program. There was no improvement in either math or Chinese when the Qinghai CAL Program II was implemented as an in-school program.

[2] On average, one-quarter of the treatment students in Qinghai CAL Program II were in schools that used regular school hours for the CAL sessions (Lai et al., 2014).

While the five studies by themselves offer interesting insights into the effectiveness of CAL sessions in raising the educational performance of rural students in China, we believe that pooling the data together can provide additional insights. Results from a pooled study will offer more external validity and statistical power. The increased power will allow for more accurate identification of heterogeneous effects and for more robustness when executing multiple hypothesis tests.[3]

Sampling, data and methods

In this section, we describe the aggregated dataset from the five CAL programs. While minor differences exist from study to study we highlight the similarities by describing the sampling and assignment of treatment and control groups, data collection, interventions, and analytical methods.

Sampling and Random Assignment

In this subsection, we summarize in four steps the sampling strategies and the randomization in each of the five CAL programs as well as present the results of statistical tests that examine a.) the *balance* of the pooled dataset; and b.) how attrition affects the balance. We first present how each program obtained the sampling frame of schools and how the sample schools were chosen. Second, we describe how we randomized the sample into treatment and control groups in each program. Third, we conduct the balance tests of randomization on the aggregated data set that we created by

[3] Our power calculations suggested that the pooled CAL study has a power of 90 percent to detect an effect size of 0.2 standard deviations of a program impact at the one percent significance level. We assumed a pre- and post-intervention correlation of 0.6 and intra-cluster correlation of 0.1. Using the Bonferroni method, our significance level for detecting the heterogeneous effects of 0.2 standard deviations is 2 percent.

pooling the individual data sets from the five programs. Fourth and finally, using the pooled data set, we also check whether the overall rate and nature of attrition are the same between the treatment and control groups.

Choosing the sample for each program consisted of several steps. The first step was to create a sampling frame. For the Migrant CAL Program, we obtained a complete list of all the migrant schools in Beijing. We then chose three districts with a high density of migrants and migrant schools. There were a total of 43 migrant schools in the three districts of Beijing. For the Shaanxi CAL Programs I and II, we chose Ankang Prefecture, one of the poorest mountainous areas in the southern region of Shaanxi Province (CNBS, 2011). Within the prefecture, we randomly selected four counties out of ten counties as our sample counties. All of the counties were nationally-designated poverty counties. We then obtained a list of all rural primary schools that had six grades. In total there were 72 schools in the sampling frame. For the Qinghai CAL Programs I and II, we chose Haidong Prefecture, which is among the poorest regions of China (CNBS, 2011). Within Haidong Prefecture, we chose the three minority autonomous counties which met our criteria of being poor and rural (Fang Lai et al. 2012) and created a sampling frame with 70 primary schools.

After creating the sampling frame, we had to choose the schools that would be in our sample. In each case, we randomly chose enough schools from the sample frame that the power of our statistical analysis allowed for at least an 80 percent chance of discovering a 0.15 standard deviation effect of the CAL program. In the Migrant CAL Program, we randomly chose 24 schools out of 43 schools for the experiment (Lai et al. 2011) encompassing 2224 grade 3 students. For the Shaanxi CAL Programs I and II, all

72 schools were included in our sample (Table 1, row 2), encompassing 2739 grade 3 and grade 5 students for Shaanxi CAL Program I and 8401 grade 3 to 6 students (Table 1, row 3) for Shaanxi CAL Program II. In the Qinghai CAL Programs I and II, we randomly chose 60 out of 70 schools (Lai et al., 2012),[4] encompassing 1828 grade 3 students for Qinghai CAL Program I and 1705 grade 3 students for Qinghai CAL Program II (Table 1, row 5).

After choosing the sample schools in each of the programs, we randomly selected the treatment and control groups. Among the 24 schools in the Migrant CAL Program, one class in each school was chosen as the treatment class and the other was taken as the control class.[5] In both of the Shaanxi CAL Programs, 36 of 72 sample schools were randomly chosen as the treatment schools and the remaining 36 schools served as control schools. Similarly, in both of the Qinghai CAL Programs 57 sample schools were randomly chosen as the treatment schools and the remaining 31 served as control schools.[6] In all treatment schools, all of the sample students were required to take the CAL sessions.

Data

The data collection approach and the survey instruments were virtually the same for all five programs. For each we conducted a baseline survey at the beginning of the study before implementation of the CAL treatment and an evaluation survey at the end of

[4] Three of the 60 schools were shut down before the program implementation. Therefore, we had a total number of 57 sample schools in the Qinghai CAL Program I (Table 1, row 4)

[5] In Lai et al. (2011), the researchers tested for spillovers by including randomly-chosen, pure control schools. In such schools there were no treatment classes. By comparing the pure control schools with the control classes in the treatment schools, it was confirmed that there were no spillovers from the treatment classes to the control classes within the same schools.

[6] In Qinghai, due to our limited supply of computers, we were only able to implement CAL in 26 schools.

each program. During each survey trained enumerators administered a standardized math test and a standardized Chinese test. Students were required to finish the tests in each subject within 25 minutes. Besides the math and Chinese tests, enumerators also collected data on the characteristics of students and their families.

Because all of the surveys were identical, we are able to create demographic and socioeconomic variables for all observations in all studies. In the current pooled study, we include variables for each student's *gender*; if the student is an *only child*; if the student has *ever used a computer* (before the CAL program); if the student's *father is illiterate*; if the student's *mother is illiterate*; whether *at least one parent has an off-farm job*, if the student has *ever used internet*; how much the student *like(s) schooling*;[7] and student *self-efficacy.*[8] A detailed summary of all the socioeconomic variables listed above is presented in Appendix 1.

When pooling the samples together, balance tests confirm that the randomization generated balanced treatment and control groups. At the time of baseline there were no significant differences in the student and parental characteristics between the treatment and control groups in the pooled sample (Table 3, column 2).

Although at baseline there was a total of 16856 students in the five CAL programs, there was an overall attrition rate of 8.5% (Table 2). In general, students

[7] To create the indicator for student's attitudes towards schooling, students were asked to rate their attitudes towards school on a 0-100 scale, where "0" indicates "extremely hates school", and "10" indicates "extremely enjoys school."

[8] The construct of Perceived Self-efficacy reflects an optimistic self-belief (Schwarzer and Jerusalem 1995). Perceived self-efficacy is an operative construct, i.e., it is related to subsequent behavior and, therefore, is relevant for clinical practice and behavior change. Jerusalem and Schwarzer developed the General Self-Efficacy Scale (GSE) in 1979, which was then widely employed in measuring self-efficacy. GSE has ten items. Each item refers to successful coping and implies an internal-stable attribution of success. In our study, we adopted the Chinese adaption of the GSE developed in (Zhang and Schwarzer 1995). In the analysis, we standardized the self-efficacy scores.

attrited because they were present during the baseline but absent (or had transferred out) during the evaluation. We do not believe that attrition affects our analysis. In the pooled data set, the attrition rates do not differ between the treatment and the control groups (Table 3, column 2). Table 1 also shows that the treatment and the control groups attrited at similar rates in each of the individual CAL programs. For example, the treatment group attrited at a rate of 6.7% and control group attrited at a rate of 6.9% in the Migrant CAL Program (row 1). In fact, if we systematically examine attrition across treatment and control groups in each of the CAL programs, no statistically significant difference between them is apparent (Table 3, column 3).[9]

In sum, at the time of the baseline of the five CAL studies, there were 16856 students in the sample. After randomly assigning classes/schools to treatment and control, there were 7584 treatment students and 9313 control students. By the end of the study 15421 students remained in the analytical sample. Of the total number of students in the sample, 6919 were treatment students and 8502 were control students.

Intervention

During each of the five programs students in the treatment groups were required to attend two 40-min CAL sessions per week in math and/or Chinese. The CAL sessions were mandatory and attendance was recorded by a teacher-supervisor. For the Migrant CAL Program and the Shaanxi CAL Program I, students in the treatment group were required to have two 40-min math CAL sessions per week. The subject was math for

[9] In order to test how attrition may have affected our ATE estimate of the program impact, we have calculated the upper and lower bounds of the ATE estimate using the method proposed by Lee (2009). Using this method, we estimate that the lower bound ATE estimate of the CAL treatment effect (in math or Chinese) on a combined test score (math + Chinese) is 0.095 SD and the upper bound is 0.103 SD. The bounded values are close to the estimated ATE (0.10 SD). Hence, the Lee Bounds analysis also confirms that attrition is not a concern in obtaining an accurate ATE in our study.

Migrant CAL Program and Shaanxi CAL Program I. The subject was Chinese for the Qinghai CAL Program I. In the Shaanxi CAL Program II and the Qinghai CAL Program II students had CAL sessions for both math and Chinese.

During all of the CAL sessions, two students shared one computer and played games that were related to either math or Chinese. The software used in CAL sessions was made up of a series of game-based learning units. The units combined animated videos (explaining the subject) with quizzes. The programs gave students feedback if they missed the questions. The CAL software was designed explicitly to provide remedial tutoring in basic competencies included in the National Uniform math and Chinese curricula. The content was exactly the same for all students within the same grade across treatment schools.

During the CAL classes, if the students had a course–related question, they were encouraged to discuss it with their teammate (the student with whom they shared the computer). The students were not allowed to discuss their questions with other teams or with the teacher-supervisor. The protocol required that the teachers could only help students with scheduling, computer hardware issues and software operations. In fact, according to our observations, the sessions were so intense that the students were almost always exclusively focused on their computers. There was little communication among the groups or between any of the groups and the teacher-supervisor. The CAL software had enough content and exercise games to cover the math/Chinese course materials for the entire experiment period and the material for each subject was sufficient to provide 80 minutes of remedial tutoring per week.

Statistical methods

Researchers use meta-analysis techniques to synthesize the results from a series of experiments, often because they do not have access to the detailed data for each study (Blettner et al. 1999). When detailed data are available, pooling data of different studies can provide improved and less-biased point estimates and afford more statistical power than performing a meta-analysis (Taioli and Bonassi 2002). Furthermore, pooling data can realize more interaction and sub-group analysis to evaluate heterogeneity. As we have the complete datasets from all five CAL experiments, we pooled the data to perform the analysis to investigate the average and heterogeneous effects of CAL.

A major objective of meta-analysis is to summarize the overall (or "combined") effect of a particular intervention across multiple studies (Hedges et al., 2009). The overall effect can be summarized using what is known in the meta-analysis literature as either a "fixed effects model" or a "random effects model." Each model makes different assumptions about the studies that are included in the meta-analysis. The different assumptions lead to different definitions of the overall effect. They also lead to different ways of using weights to estimate the overall effect. In the meta-analysis conducted in our paper, we do not seek to make rigid assumptions about the underlying true effect(s). We therefore use both fixed effects and random effects models. We find that our results are substantively similar and robust across models.

Under the fixed effect model, the researcher assumes that there is a single overall effect size (a "true" effect size) of the intervention that is being analyzed across multiple studies. Each study has information that can be used to estimate this single overall effect size. Studies that provide more information for estimating the overall effect size (for example, some studies measure effects with greater precision than other studies) are

assigned larger weights than studies that provide less information. Importantly, the only source of error in estimating the overall effect size in the fixed effects model is the random error (the lack of information with which to estimate effects) within studies.

By contrast, under the random effects model, the researcher assumes (a) that there is a distribution of true effect sizes of the intervention (for example, the intervention may have a larger impact in some contexts or with some populations as compared to others) and (b) that the studies included in the meta-analysis are a random sample of the distribution of true effect sizes of the intervention. Thus in the random effects model, the researcher estimates the mean of this distribution of true effects rather than a single overall true effect as in the fixed effects model. When estimating the mean of this distribution, the random effects model accounts for two possible sources of error (rather than the single source of within-study error as in the fixed effects model). First, each study is used to estimate the true effect for a specific context (or for a specific population). Second, the true effects for specific contexts are used to estimate the mean of the distribution of true effects. The combined mean effect therefore depends not only on the precision of each study (the degree of within-study error as in the fixed effects model) but also on the number of studies included in the meta-analysis.

It should be noted that, similar to the fixed effects model, the random effects model also places greater weight on studies that provided greater information for estimating true effect sizes. However, in the random effects model each study is estimating a different true effect size (drawn from a distribution of true effect sizes). To account for this difference, the weights assigned under the random effects model are more balanced than the weights assigned under the fixed effects model. In other words,

studies that estimate effects with greater precision are less likely to dominate the estimation of the total effect in the random effects model (and studies that estimate effects with less precision are less likely to be discounted) compared with the fixed effects model.

Inside the framework of both our fixed effects and random effects approaches, we also estimate both unadjusted and adjusted ordinary least squares (OLS) regression models. The unadjusted analysis regresses the outcome variable (i.e. standardized math and Chinese test scores) on a dummy variable that measures treatment status (CAL intervention). While no other control variables are included in the unadjusted analysis, we do hold constant a pre-program outcome variable (i.e., the baseline math and/or Chinese test score). In summary, then, the unadjusted model that we estimate is:

$$y_{is} = \alpha + \beta * treatment_s + \theta * y_{0is} + \varepsilon_{is} \qquad (1)$$

where y_{is} is the outcome variable after the CAL program for student i in school s; $treatment_s$ is a dummy variable measuring treatment status (equal to one for students in the CAL treatment group and zero otherwise) and ε_{is} is a random disturbance term clustered at the school level.[10] We also control for y_{0is}, the baseline math test score and/or Chinese test score for student i in school s.

The model in the adjusted analysis is the same as the unadjusted analysis, but, we also include a series of control variables to improve statistical efficiency. The adjusted model that we estimate is:

$$y_{is} = \alpha + \beta * treatment_s + \theta * y_{0is} + X_{is}. + \varepsilon_{is} \qquad (2)$$

[10] Following Cameron, Gelbach, and Miller (2011) we correct for the highest level of clustering. In our case, it is the school level.

where all notation is the same as in the unadjusted model (equation 1), except we also include a set of control variables, X_{is}. Specifically, X_{is} is a vector of student demographic and socioeconomic variables (*gender*; *only child*; *ever used a computer*; *father is illiterate*; *mother is illiterate*; *at least one parent has an off-farm job*; *ever used internet*; *like schooling*; and *self-efficacy*). These variables are all generated using the baseline data.

By construction, in both models the coefficient of the dummy variable $treatment_s$, β, is equal to the unconditional difference in the outcome (y_{is}-y_{0is}) between the treatment and control groups over the program period. In other words, β measures how the treatment group changed in the standardized math/Chinese test score levels after the CAL program relative to the control group. In summary, in the results section below, we report the results of our analysis from estimating Equation (1) with control variables (the adjusted model) and without control variables (the unadjusted model) using both fixed effect and random effects models.

Results

Our analysis using the pooled data set shows that the CAL treatment in math or Chinese significantly improves the student test scores of the treatment group relative to the control group (Table 4).11 The CAL treatment in math or Chinese is found to improve the total test scores by 0.10 standard deviations (significant at the 1% level,

[11] In the results section of the paper, when we use the term the CAL treatment in math or Chinese, we mean either of the CAL programs—that is, either the math CAL program or the Chinese CAL program.

row 1, columns 1 to 4).[12] The estimates of the impact remain the same whether we use the adjusted, unadjusted, fixed effect or random effects model.[13]

While there is a significant overall effect, we find that the program impact varies when we implement different types of CAL treatment (Table 5). When we use the CAL treatment that provides remedial tutoring for math only, math test scores rise by 0.11 standard deviations (significant at the 1% level, row 1, columns 1 and 2). The CAL treatment in math alone did not have any spillover effects on Chinese test scores (Table 4, row 1, columns 3 and 4). In contrast, the CAL treatment in Chinese only had a large positive impact on Chinese test scores which rose by 0.17 to 0.18 standard deviations (significant at the 1% or 5% level, row 2, columns 3 and 4). Importantly, when we ran the CAL treatment in Chinese only, we also observed a positive and significant spillover onto math test scores (of 0.25 standard deviations—significant at the 1% levels, row 2, columns 1 and 2).

The results of our study also show that some of the CAL programs created impacts that extend beyond test score effects. Student self-efficacy improved if students attended the CAL program in math only (Table 6, row 2, columns 3 and 4). Such CAL treatments improved student self-efficacy by 0.08 standard deviations (significant at the 10% level, row 2, columns 3 and 4). However, there was no impact on the students who received CAL treatment in Chinese only (row 3, columns 3 and 4). The above results

[12] In the rest of the paper, when we use the term *total test scores* we mean the sum of math and Chinese test scores. Recall that in all CAL programs (whether we treated students with the CAL math program by itself or with the CAL Chinese program by itself or with both the CAL math and Chinese programs), we gave students two standardized tests (one in math and one in Chinese).

[13] As a robustness check, we tested the program impact of CAL treatment in math or Chinese by including program dummies. The estimated program impact remains the same when we compare the specification without program dummies with the specification with program dummies. The estimation results are available upon request.

hold true under both the fixed effects and random effects models. . One of the reasons that the math CAL was able to make an impact on self-efficacy may be that practices in math may involve more of a problem-solving process that can boost student self-efficacy. In contrast, language exercises mainly enforce the memory of vocabularies and grammar and understanding of sentences or paragraphs, which may be less likely to increase student self-evaluation of their capacity to accomplish learning tasks.

The analysis shows that how CAL is implemented also matters. Specifically, our results suggest that out-of-school CAL programs seem to work better than in-school CAL programs. Using our pooled data set and either the fixed effects or the random effects model), the out-of school CAL treatment had a larger positive impact on student total test scores (that is, math + Chinese scores) than the in-school CAL treatment. The out-of-school CAL program had an impact (0.15 standard deviations—Table 7, row 2, columns 1 and 2) that was higher than the in-school CAL program (0.03 standard deviations—Table 7, row 1, columns 1 and 2). Importantly, the gap between the two programs (0.12 standard deviations or 0.15 – 0.03) is significant at the 1% level (Table 7, rows 1 and 2, columns 1 and 2). The difference in the program impacts on the total test score (math + Chinese scores) is mainly driven by the differences in the program impacts on math scores. The gap in the math test scores from the out-of-school (row 2, columns 3 and 4) and the in-school CAL programs (row 1, columns 3 and 4) is 0.19 standard deviations (0.23 – 0.04 using the fixed effect model) or 0.18 standard deviations (0.23-0.05 using the random effects model). This difference is significant at the 1% level. Neither program had a significant impact on Chinese test scores. Moreover, the gap between the impacts of the two types of programs on Chinese test scores is small (0.07-0.01=0.06 using the

fixed effect model or 0.06-0.01=0.05 using the random effects model) and is insignificant.

Our results indicate that the out-of-school program was more effective than the in-school program. The in-school program had a much smaller impact on student academic performance than the out-of-school program, which is consistent with Lai's study (2014). While we do not know for sure, the reason for the absence of an in-school effect may be that in-school programs substituted effective teaching and cancelled out the positive impact of the CAL classes.

The pooled analysis also identified systematic differences in CAL heterogeneous program effects. According to our analysis, boys gained more in Chinese test scores than girls from the Chinese only CAL treatment (Table 8). More specifically, girls gained 0.12 standard deviations in Chinese (and this coefficient was insignificant at the 10% level, row 4, columns 3 and 4) while boys gained 0.23-0.24 standard deviations (0.12+0.11 in the fixed effect model or 0.12+0.12 in the random effects model; significant at the 5% level, rows 2 and 4, columns 3 and 4). This suggests that, using the fixed effect or the random effects model, the gap in Chinese test scores between boys and girls is 0.11 or 0.12 standard deviations (indicated by the coefficient on the interaction term, row 2, columns 3 and 4). This is significant at the 10% level in the fixed effect model and 5% level in the random effects model. In contrast, we do not find heterogeneous effects in math test scores between the girls and boys when the math only treatment was implemented (the coefficient on the interaction term between CAL treatment in math only and the gender dummy is insignificant, row 1, columns 1 and 2). In other words,

girls and boys benefit similarly from the math only CAL treatment no matter which model is used.[14]

One possible reason that boys gain more from a CAL treatment in Chinese is that boys had lower levels of Chinese than girls before the program. Since the content of the software only covered the course material and provided remedial tutoring to the students, it may have been more useful to students with lower levels of knowledge than to students with higher levels of learning. By looking at the baseline level of Chinese of girls and boys and controlling for school fixed effects, we find that boys scored 0.17 SD lower than girls in Chinese (significant at 1% level). Other studies have also suggested that a remedial program tends to help the poorer performing students more than the better performing students (Banerjee et al., 2007).

Despite having a large sample and high power, we do not find significant heterogeneous effects by student initial academic achievement (Table 9). For the math-only CAL treatment, better performing students (those scoring in the top 50 percentile at the baseline) gained as much as those scoring in the bottom 50 percentile at the baseline (the coefficient on the interaction term is insignificant, rows 1, columns 1-2). Although the coefficients in the fixed effects model (0.09) and the random effects model (0.08) suggest that there might be heterogeneous effects of the Chinese-only CAL treatment on student Chinese test scores, the coefficient on the interaction term between the treatment variable and the indicator for bottom 50% student in baseline Chinese test is not

[14] We have also conducted a robustness check by dividing the sample into boys and girls and estimating the program impact among each gender subgroup. The results are consistent with those that use an interaction term between the treatment variable and the gender dummy (Table 8). The estimation results of the robustness check are available upon request.

significant (row 2, columns 3 and 4). Therefore, we cannot reject the hypothesis that there are no heterogeneous effects of Chinese only CAL on Chinese test scores.

Conclusion

In this paper we present the results from a pooled dataset of five randomized field experiment of CAL programs in rural China. The combined studies include 15421 primary school students. In total, there are 6919 students in the treatment group and 8502 in the control group. Students in the treatment arm received two 40-minute CAL sessions per subject per week, during which, students played computer-based games that required them to practice using their knowledge of math and/or Chinese.

Our results suggest that overall the CAL program has a robust and consistently positive impact on student academic performance as measured by standardized test scores. The additional drills and exercise provided by the CAL software, the freshness of the novel technology and the prompt interaction and immediate feedback from computers may have all contributed to the positive impact in student learning. The impacts of specific programs ranged from 0.11 to 0.25 standard deviations in math test scores and 0.03 to 0.18 standard deviations on Chinese test scores. The data also suggest that there are spillover effects of Chinese CAL programs on math test scores. The Chinese-only program improved student Chinese test score by 0.25 standard deviations.

The study also finds that student self-efficacy improved by 0.08 standard deviations when students were treated by our CAL math programs. However, there are no effects on student self-efficacy when students had Chinese CAL sessions. One of the reasons that the math CAL was able to make an impact on self-efficacy may be that

practices in math may involve more of a problem-solving process that can boost student self-efficacy. In contrast, language exercises mainly enforce the memory of vocabularies and grammar and understanding of sentences or paragraphs, which may be less likely to increase student self-evaluation of their capacity to accomplish learning tasks.

Our results indicate that the out-of-school program was more effective than the in-school program and that boys benefited more than girls from CAL treatment in Chinese. The in-school program had a much smaller impact on student academic performance than the out-of-school program, which is consistent with Lai's study (2014). While we do not know for sure, the reason for the absence of an in-school effect may be that in-school programs substituted effective teaching and cancelled out the positive impact of the CAL classes. We also found that boys gained more in Chinese test scores than girls in CAL treatment in Chinese. Boys gained 0.11 to 0.12 standard deviations more than the girls from the CAL treatment in Chinese only.

Many questions are worth exploring in future studies. More studies need to be conducted to investigate the mechanisms through which the CAL program improves student achievement. Is it because the program is better at adjusting to the pace of learning of the individual than regular teaching? Is it due to the more complete and immediate feedback of student performance that helped the students? Is it because the pairs of the students discussed and collaborated in CAL classes that made learning more efficient? Or is it because the use of software boosted the students' motivation to learn in general? The answers to these questions have important implications for increasing the effectiveness of the CAL programs and improving teacher practices in regular classes.

In summary, our results suggest that CAL is an effective and cost-effective solution to bridging the educational gap between the rural and urban students in China. Previous studies suggest there is a significant educational gap between the rural and urban students (Fu and Ren 2010). CAL is a potential solution to narrowing the gap if it is effective in improving the academic achievement of the rural students. It is also cost-effective, given that the government is committed to building computer labs in all rural schools. Computer hardware itself is already a sunk cost as it has been part of the government's Twelfth Five-Year Plan. The marginal costs that are needed to execute the program include teacher training, administration costs and allowance for CAL teacher-supervisors. Using the method suggested by (Dhaliwal et al. 2011), we calculate the cost per unit of improvement in student learning to be 24 USD/SD.[15] The cost-effectiveness of our program is comparable to the CAL program conducted in India (Banerjee et al., 2007).

However, attention is needed regarding the implementation strategy of the CAL program. For example, our results suggest that the program is more effective if it is implemented during a less productive period of time for schooling (e.g. out-of-school program) than replacing teacher instruction in the regular classes (e.g. in-school program). We designed and implemented the CAL protocol in a way that made it easy

[15] We calculate the total annual cost of the program to 16,100 USD (in 2014, after taking inflation into account). We then divide the total cost by total impact (total impact=average program effect multiplied by the total number of students attending CAL sessions): 16,100 USD/(0.10 SD * 6714 students)=24.0 USD/SD. According to the estimates provided by (Banerjee et al. 2007), the CAL program in India costs 21.4 USD/SD (in 2002) and 28.2 USD/SD (in 2014)—also excluding the costs of computers.

and attractive for teachers to follow. We conducted an intensive teacher training where teachers learned about the protocol and practiced using the software. We also provided subsidies to compensate teacher-supervisors for any additional workload associated with the CAL program. To ensure that principals do not shirk on the implementation, it may be helpful for authorities to incentivize them by "contracting" or linking program outcomes with an evaluation of overall performance or taking advantage of certain forms of payment conditional on program implementation.

References

Bakar, Aysegul, Yavuz Inal, Kursat Cagiltay, Aysegul Bakar, Yavuz Inal, and Kursat Cagiltay. 2006. "Use of Commercial Games for Educational Purposes: Will Today's Teacher Candidates Use Them in the Future?" In , 2006:1757–62. http://www.editlib.org/p/23243/.

Baker, Scott, Russell Gersten, and Dae-Sik Lee. 2002. "A Synthesis of Empirical Research on Teaching Mathematics to Low-Achieving Students." *The Elementary School Journal*, 51–73.

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): 1235–64. doi:10.1162/qjec.122.3.1235.

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. doi:10.1257/pol.1.1.52.

Blettner, Maria, Willi Sauerbrei, Brigitte Schlehofer, Thomas Scheuchenpflug, and Christine Friedenreich. 1999. "Traditional Reviews, Meta-Analyses and Pooled Analyses in Epidemiology." *International Journal of Epidemiology* 28 (1): 1–9.

Borenstein, Michael, Larry V. Hedges, Julian PT Higgins, and Hannah R. Rothstein. 2011. *Introduction to Meta-Analysis*. John Wiley & Sons. http://books.google.com/books?hl=en&lr=&id=JQg9jdrq26wC&oi=fnd&pg=PT14&dq=Borenstein+2007+introduction+to+meta-analysis&ots=VHZ0KMqycy&sig=E8Z47equhB2f43NUIAw3tgFgRyY.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2). http://amstat.tandfonline.com/doi/abs/10.1198/jbes.2010.07136.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2011. *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education*. Cambridge, MA: MIT Press. http://books.google.com/books?hl=en&lr=&id=uSywAQAAQBAJ&oi=fnd&pg=PA285&dq=Dhaliwal+cost+effectiveness&ots=4i3x6cArpy&sig=4o0Wn7NKdj4PW_HgdkCbQ_AySxk.

Ebner, Martin, and Andreas Holzinger. 2007. "Successful Implementation of User-Centered Game Based Learning in Higher Education: An Example from Civil Engineering." *Computers & Education* 49 (3): 873–90. doi:10.1016/j.compedu.2005.11.026.

2010 Educational Inequality under China's Rural–urban Divide: The Hukou System and Return to Education. Environment and Planning. A 42(3): 592.

Gorjian, Bahman, Seyyed Rahim Moosavinia, Kamal Ebrahimi Kavari, Parviz Asgari, and Alireza Hydarei. 2011. "The Impact of Asynchronous Computer-Assisted Language Learning Approaches on English as a Foreign Language High and Low Achievers' Vocabulary Retention and Recall." *Computer Assisted Language Learning* 24 (5): 383–91.

Hannum, Emily. 1999. "Poverty and Basic-Level Schooling in China: Equity Issues in the 1990s." *Prospects* 29 (4): 561–77.

Hativa, Nira. 1988. "Computer-Based Drill and Practice in Arithmetic: Widening the Gap between High-and Low-Achieving Students." *American Educational Research Journal* 25 (3): 366–97.

Hyland, Ken. 1993. "ESL Computer Writers: What Can We Do to Help?" *System* 21 (1): 21–30.

Lai, F., R. Luo, L. Zhang, X. Huang, and S. Rozelle. 2011. *Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing*. REAP working paper.

Lai, Fang, Linxiu Zhang, Qinghe Qu, Xiao Hu, Yaojiang Shi, Matthew Boswell, and Scott Rozelle. 2012. *Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Public Schools in Rural Minority Areas in Qinghai, China*. REAP Working Paper 237. http://fsi.fsi.stanford.edu/sites/default/files/CAL_qinghai_aug16_mwb_fang_edcc_clean_sdr.pdf.

Liu, Min, Peggy Hsieh, YoonJung Cho, and Diane Schallert. 2006. "Middle School Students' Self-Efficacy, Attitudes, and Achievement in a Computer-Enhanced Problem-Based Learning Environment." *Journal of Interactive Learning Research* 17 (3): 225–42.

Maag, Margaret. 2004. "The Effectiveness of an Interactive Multimedia Learning Tool on Nursing Students' Math Knowledge and Self-Efficacy." *Computers Informatics Nursing* 22 (1): 26–33.

Madorin, Sandra, and Carroll Iwasiw. 1999. "The Effects of Computer-Assisted Instruction on the Self-Efficacy of Baccalaureate Nursing Students." *The Journal of Nursing* Education 38 (6): 282−85.

Malamud, O. and C. Pop-Eleches, 2011. Home Computer Use and the Development of Human Capital. *The Quarterly Journal of Economics*: 126(2), 987-1027.

Mo, Di, Hongmei Yi, Linxiu Zhang, Yaojiang Shi, Scott Rozelle, and Alexis Medina. 2012. "Transfer Paths and Academic Performance: The Primary School Merger Program in China." *International Journal of Educational Development* 32 (3): 423–31.

Newell, Marie-Louise, Hoosen Coovadia, Marjo Cortina-Borja, Nigel Rollins, Philippe Gaillard, and Francois Dabis. 2004. "Mortality of Infected and Uninfected Infants Born to HIV-Infected Mothers in Africa: A Pooled Analysis." *The Lancet* 364 (9441): 1236–43.

Ong, Chorng-Shyong, and Jung-Yu Lai. 2006. "Gender Differences in Perceptions and Relationships among Dominants of E-Learning Acceptance." *Computers in Human Behavior* 22 (5): 816–29.

Osín, Luis. 1998. *Computers in Education in Developing Countries: Why and How?* Education and Technology Team, Human Development Network, World Bank. http://www.tigweb.org/action-tools/projects/download/5949/v3n1.pdf.

Pal, Joyojeet, Udai Singh Pawar, Eric A. Brewer, and Kentaro Toyama. 2006. "The Case for Multi-User Design for Computer Aided Learning in Developing Regions." In *Proceedings of the 15th International Conference on World Wide Web*, 781–89. ACM. http://dl.acm.org/citation.cfm?id=1135896.

Prinsen, Fleur Ruth, M. L. L. Volman, and Jan Terwel. 2007. "Gender-Related Differences in Computer-Mediated Communication and Computer-Supported Collaborative Learning." *Journal of Computer Assisted Learning* 23 (5): 393–409.

Schaefer, Scott, and Joe Warren. 2004. "Teaching Computer Game Design and Construction." *Computer-Aided Design* 36 (14): 1501–10.

Schwarzer, Ralf, and Matthias Jerusalem. 1995. "Generalized Self-Efficacy Scale." *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs* 1: 35–37.

Taioli, Emanuela, and Stefano Bonassi. 2002. "Methodological Issues in Pooled Analysis of Biomarker Studies." *Mutation Research/Reviews in Mutation Research* 512 (1): 85–92.

Tüzün, Hakan, Meryem Yılmaz-Soylu, Türkan Karakuş, Yavuz İnal, and Gonca Kızılkaya. 2009. "The Effects of Computer Games on Primary School Students' Achievement and Motivation in Geography Learning." *Computers & Education* 52 (1): 68–77.

Vichitvejpaisal, Phongthara, Sukalya Sitthikongsak, Benjamas Preechakoon, Kanita Kraiprasit, Sudta Parakkamodom, Chitprapa Manon, and Suppat Petcharatana. 2001. "Does Computer-Assisted Instruction Really Help to Improve the Learning Process?" *Medical Education* 35 (10): 983–89.

Vogel, Jennifer J., David S. Vogel, Jan Cannon-Bowers, Clint A. Bowers, Kathryn Muse, and Michelle Wright. 2006. "Computer Gaming and Interactive Simulations for Learning: A Meta-Analysis." *Journal of Educational Computing Research* 34 (3): 229–43.

Watkins, David. 2000. “Learning and Teaching: A Cross-Cultural Perspective.” *School Leadership & Management* 20 (2): 161–73.

Zhang, Jian Xin, and Ralf Schwarzer. 1995. “Measuring Optimistic Self-Beliefs: A Chinese Adaptation of the General Self-Efficacy Scale.” *Psychologia: An International Journal of Psychology in the Orient*. http://psycnet.apa.org/psycinfo/1996-35921-001.

Table 1. An overview of the five CAL programs

	CAL program	Location	Subject	Duration	Treatment group	Number of treatment students	Treatment attrition rate	Control group	Number of control students	Control attrition rate
(1)	Migrant CAL Program	Beijing	Math	One semester	24 classes	943	6.7%	24 classes	1281	6.9%
(2)	Shaanxi CAL Program I	Shaanxi	Math	One semester	36 schools	1277	2.0%	36 schools	1462	1.4%
(3)	Shaanxi CAL Program II	Shaanxi	Math and Chinese	Two semesters	36 schools	3912	9.6%	36 schools	4489	10.8%
(4)	Qinghai CAL Program I	Qinghai	Chinese	One semester	26 schools	737	10.9%	31 schools	1091	7.1%
(5)	Qinghai CAL Program II	Qinghai	Math and Chinese	Two semesters	26 schools	715	17.1%	31 schools	990	14.3%

Table 2. Student's attrition status across the five CAL programs and the whole sample

Dependent variable: student's attrition status (1=attrited;0=otherwise)						
	Migrant CAL Program	Shaanxi CAL Program I	Qinghai CAL Program I	Shaanxi CAL Program II	Qinghai CAL Program II	All five programs in columns (1)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)
[1] CAL treatment in math or Chinese (1=yes; 0=no)	-0.01	-0.02	0.06	-0.01	0.02	-0.00
	(0.01)	(0.03)	(0.04)	(0.02)	(0.03)	(0.02)
[2] Observations	2,197	2,739	1,819	8,400	1,701	16,856
[3] R-squared	0.001	0.002	0.006	0.000	0.001	0.000

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The test aims to show whether attrition rates are different between the treatment and control groups in each CAL program and five programs all together. The tests regress the attrition status (1=attrited student; 0=remaining student) on the indicator of CAL treatment (1=yes; 0=no) for each program and all five programs.

Table 3. Ordinary least squares analysis of the differences in student's characteristics between the attrited students and non-attrited students, and between the treatment and control students before and after attrition

	Differences between attrited students and non-attrited students	Differences between treatment students and control students **before attrition**	Differences between treatment students and control students **after attrition**
	(1)	(2)	(3)
[1] Standardized baseline math test score (standard deviations) [a]	-0.19***	0.01	-0.00
	(0.04)	(0.08)	(0.08)
[2] Standardized baseline Chinese test score (standard deviations) [b]	-0.22***	0.02	0.01
	(0.04)	(0.07)	(0.07)
[3] Standardized baseline total test score (math + Chinese, standard deviations) [c]	-0.23***	0.01	0.01
	(0.04)	(0.09)	(0.09)
[4] Gender (1=male; 0=female)	0.02*	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
[5] Only child (1=yes; 0=no)	-0.01	-0.00	-0.00
	(0.01)	(0.03)	(0.03)
[6] Ever used a computer (1=yes; 0=no)	-0.10***	0.02	0.01
	(0.03)	(0.05)	(0.05)
[7] Father is illiterate (1=yes; 0=no)	-0.00	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
[8] Mother is illiterate (1=yes; 0=no)	-0.00	-0.01	-0.02
	(0.03)	(0.03)	(0.03)
[9] At least one parent has an off-farm job (1=yes; 0=no)	-0.04*	-0.02	-0.02
	(0.02)	(0.03)	(0.03)
[10] Ever used internet (1=yes; 0=no)	-0.02	0.01	0.01
	(0.02)	(0.03)	(0.03)
[11] Like school (1-100 points)	-1.06*	-0.35	-0.42
	(0.57)	(0.70)	(0.73)
[12] Baseline self efficacy (standard deviations)	-0.04**	-0.01	-0.01
	(0.02)	(0.02)	(0.04)
[13] Observations	16,856	16,856	15,421

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The test in column (1) aims to show who are more likely to be attrited from the sample. The tests in columns (2) and (3) aim to show whether the characteristics of the treatment and control groups are balanced before and after attrition.

Column (1) regress the attrition status on student characteristics (variables in Appendix 1). The tests in column (2) and (3) regress the student characteristics (variables in Appendix 1) on the treatment status one at a time.

[ab] The Standardized baseline math/Chinese score is the normalized math/Chinese score on the math/Chinese test that is given to all sample students before CAL programs.
[c] To generate a standardized baseline total score, we first standardized the math and Chinese scores separately and then added them together to gain the total standardized scores.

Table 4. Ordinary least squares analysis of the impact of CAL program on student's total score

Dependent variable: standardized evaluation total test score (math + Chinese)	Fixed effect		Random effects	
	Without control (1)	With control (2)	Without control (3)	With control (4)
[1] CAL treatment in math or Chinese (1=yes; 0=no)	0.10***	0.10***	0.10***	0.10***
	(0.03)	(0.03)	(0.03)	(0.03)
[2] Standardized baseline total test score (math + Chinese) [a]	0.68***	0.67***	0.68***	0.67***
	(0.01)	(0.01)	(0.01)	(0.01)
[3] Gender (1=male; 0=female)		-0.00		-0.01
		(0.01)		(0.01)
[4] Only child (1=yes; 0=no)		-0.02		-0.01
		(0.02)		(0.02)
[5] Ever used a computer (1=yes; 0=no)		0.01		0.01
		(0.02)		(0.02)
[6] Father is illiterate (1=yes; 0=no)		-0.08***		-0.08***
		(0.03)		(0.03)
[7] Mother is illiterate (1=yes; 0=no)		-0.01		-0.01
		(0.02)		(0.02)
[8] At least one parent has an off-farm job (1=yes; 0=no)		0.02		0.02
		(0.02)		(0.02)
[9] Ever used internet (1=yes; 0=no)		0.01		0.01
		(0.02)		(0.02)
[10] Like school (1-100 points)		0.00***		0.00***
		(0.00)		(0.00)
[11] Baseline self efficacy (standard deviations)		0.02**		0.02**
		(0.01)		(0.01)
[12] Constant	-0.02	-0.12***	-0.02	-0.12**
	(0.02)	(0.05)	(0.02)	(0.05)
[13] Observations	15,421	15,421	15,421	15,421
[14] R-squared	0.455	0.457	0.450	0.452

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level

Note: The test aims to show the impact of the CAL treatment in math or Chinese on student total test scores.

The tests regress the student standardized evaluation total test score (math + Chinese) on the indicator of CAL treatment in math or Chinese (1=treatment student; 0=control student). Columns (1) and (2) use the fixed effect model and columns (3) and (4) use the random effects model. All tests control for standardized baseline total test score. Columns (2) and (4) control for student characteristics that are listed in Appendix 1, rows (4)-(12).

[a] To generate a standardized baseline total score, we first standardized the math and Chinese scores separately and then added them together to gain the total standardized scores.

Table 5. Ordinary least squares analysis of the impact of different CAL programs on student test scores

Dependent variable: standardized evaluation test score (standard deviations)				
	Math score		Chinese score	
	Fixed effect (1)	Random effects (2)	Fixed effect (3)	Random effects (4)
[1] CAL treatment in math only (1=yes; 0=no)	0.11***	0.11***	0.04	0.03
	(0.04)	(0.04)	(0.03)	(0.04)
[2] CAL treatment in Chinese only (1=yes; 0=no)	0.25***	0.25***	0.18***	0.17***
	(0.06)	(0.06)	(0.06)	(0.06)
[3] Controls [a]	Yes	Yes	Yes	Yes
[4] Observations	15,421	15,421	15,421	15,421
[5] R-squared	0.329	0.326	0.381	0.338

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The tests aim to show the impact of the different CAL treatments on student math test scores and Chinese test scores.

The tests in columns (1) and (2) regress student standardized evaluation math test score on indicators of CAL treatment in math only (1=only math treatment; 0=otherwise), CAL treatment in Chinese only (1=only Chinese treatment; 0=otherwise) and CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise). Columns (3) and (4) use the student standardized evaluation Chinese test score as the outcome variable. Columns (1) and (3) use the fixed effect model and columns (2) and (4) use the random effects model. All tests control for standardized baseline test score. All tests control for student characteristics that are listed in Appendix 1, rows (4)-(12).

[a] Control variables include all variables in rows (4)-(12) in Appendix 1. The baseline test scores we control for vary with the outcome variables. If the dependent variable is standardized evaluation math test scores, then we control for standardized baseline math test score. If the dependent variable is standardized evaluation Chinese test scores, then we control for standardized baseline Chinese test score. Also, indicator for CAL treatment in both math and Chinese served as control in this analysis.

Table 6. Ordinary least squares analysis of the impact of CAL programs on student self-efficacy

Dependent variable: Evaluation self efficacy (standard deviations)	(1) Fixed effect	(2) Random effects	(3) Fixed effect	(4) Random effects
[1] CAL treatment in math or Chinese (1=yes; 0=no)	0.03	0.03		
	(0.03)	(0.03)		
[2] CAL treatment in math only (1=yes; 0=no)			0.08*	0.08*
			(0.04)	(0.04)
[3] CAL treatment in Chinese only (1=yes; 0=no)			0.04	0.04
			(0.07)	(0.07)
[4] Controls [a]	Yes	Yes	Yes	Yes
[5] Observations	15,421	15,421	15,421	15,421
[6] R-squared	0.078	0.077	0.079	0.078

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The test aims to show the effects of the different CAL treatments on student self-efficacy.

The tests in columns (1) and (2) regress the student evaluation self-efficacy score on the indicator of CAL treatment in math or Chinese (1=yes; 0=no). Columns (3) and (4) regress the student evaluation self-efficacy score on the indicators of CAL treatment in math only (1=only math treatment; 0=otherwise), CAL treatment in Chinese only (1=only Chinese treatment; 0=otherwise) and CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise). Columns (1) and (3) use the fixed effect model and columns (2) and (4) use the random effects model. All tests control for the standardized baseline self-efficacy and student characteristics that are listed in Appendix 1, rows (2)-(3) and rows (4)-(11).

[a] Control variables include all variables in rows (1)-(2) and rows (4)-(12) in Appendix 1. Also, indicator for CAL treatment in both math and Chinese served as control in columns (3) and (4).

Table 7. Ordinary least squares of the impact of out-of-school and in-school CAL program on student academic outcomes

Dependent variable: standardized evaluation test score (standard deviations)	Total score (math + Chinese)		Math score		Chinese score	
	Fixed effect (1)	Random effects (2)	Fixed effect (3)	Random effects (4)	Fixed effect (5)	Random effects (6)
[1] In-school CAL treatment in both math and Chinese (1=yes; 0=no)	0.03	0.03	0.04	0.05	0.01	0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
[2] Out-of-school CAL treatment in both math and Chinese (1=yes; 0=no)	0.15***	0.15***	0.23***	0.23***	0.07	0.06
	(0.05)	(0.05)	(0.08)	(0.08)	(0.08)	(0.09)
[3] Controls [a]	Yes	Yes	Yes	Yes	Yes	Yes
[4] Observations	15,421	15,421	15,421	15,421	15,421	15,421
[5] R-squared	0.455	0.450	0.327	0.324	0.344	0.337

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level

Note: The tests aim to show the effects of the in-school CAL treatment and out-of-school CAL treatment in both math and Chinese on student test scores.

Column (1) and (2) regress the student standardized evaluation total test score (math + Chinese) on indicators of the in-school CAL treatment in both math and Chinese (1=yes; 0=no) and the out-of-school CAL treatment in both math and Chinese (1=yes; 0=no). Column (3) and (4) use the student standardized evaluation math test score as the outcome variable. Column (5) and (6) use the student standardized evaluation Chinese test score as the outcome variable. Columns (1), (3) and (5) use the fixed effect model and columns (2), (4) and (6) use the random effect model. All tests control for student characteristics that are listed in Appendix 1, rows (4)-(12).

[a] Control variables include all variables in rows (4)-(12) in Appendix 1. The baseline test scores we control for vary with the outcome variables. If the dependent variable is standardized evaluation total test scores, then we control for standardized baseline total test score. If the dependent variable is standardized evaluation math test scores, then we control for standardized baseline math test score. If the dependent variable is standardized evaluation Chinese test scores, then we control for standardized baseline Chinese test score.

Table 8. Ordinary least squares analysis of the heterogeneous effects of CAL treatment on student test score by student gender

Dependent variable: standardized evaluation test score (standard deviations)				
	Math score		Chinese score	
	Fixed effect (1)	Random effects (2)	Fixed effect (3)	Random effects (4)
[1] CAL treatment in math only (1=yes; 0=no) * Gender (1=male; 0=female)	-0.01[b]	-0.01[b]		
	(0.04)	(0.04)		
[2] CAL treatment in Chinese only (1=yes; 0=no) * Gender (1=male; 0=female)			0.11*[c]	0.12** [c]
			(0.06)	(0.06)
[3] CAL treatment in math only (1=yes; 0=no)	0.12**	0.12**	0.03	0.03
	(0.05)	(0.05)	(0.04)	(0.04)
[4] CAL treatment in Chinese only (1=yes; 0=no)	0.25***	0.25***	0.12*	0.12*
	(0.06)	(0.06)	(0.06)	(0.06)
[5] Gender (1=male; 0=female)	0.01	0.01	-0.05***	-0.06***
	(0.02)	(0.02)	(0.02)	(0.02)
[6] Controls [a]	Yes	Yes	Yes	Yes
[7] Observations	15,421	15,421	15,421	15,421
[8] R-squared	0.329	0.326	0.345	0.338

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The tests aim to show the heterogeneous effects of different CAL treatments on test scores by student gender.

Columns (1) and (2) regress student standardized evaluation math test score on the main components and the interaction terms of student gender and the CAL treatment in math only (1=only math treatment; 0=otherwise), and the main components and the interaction of student gender and the CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise). Columns (3) and (4) regress students standardized evaluation Chinese test score on the main components and the interaction term of student gender and the indicators of CAL treatment in Chinese only (1=only Chinese treatment; 0=otherwise), and the main components and the interaction of the CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise) and student gender. Columns (1) and (3) use the fixed effect model and columns (2) and (4) use the random effects model. All tests control for standardized baseline test score. All tests controlled for student characteristics that are listed in Appendix 1, rows (4)-(12).

[a] Control variables include all variables in rows (4)-(12) in Appendix 1. The baseline test scores we control for vary with the outcome variables. If the dependent variable is standardized evaluation math test scores, then we control for standardized baseline math test score. If the dependent variable is standardized evaluation Chinese test scores, then we control for standardized baseline Chinese test score. Also, indicator for CAL treatment in both math and Chinese served as control in this analysis.

[b] To reach a significance level of 0.1, the two heterogeneous tests of the CAL treatment in math need to have a p-value of 0.05 *each* (using the Bonferroni method). In other words, the interaction term in row (1), columns (1)-(2), need to be significant at the 5% level after adjusting for multiple tests of heterogeneous effects. The results suggest that we cannot reject the hypothesis that CAL treatment in math only does not have heterogeneous effects by gender.

[c] To reach a significance level of 0.1, the two heterogeneous tests of the CAL treatment in Chinese need to have a p-value of 0.05 *each* (using the Bonferroni method). In other words, the interaction term in row (2), columns (3)-(4), need to be significant at the 5% level after adjusting for multiple tests of heterogeneous effects. The results suggest that we can reject the hypothesis that CAL treatment in Chinese only does not have heterogeneous effects by gender.

Table 9. Ordinary least squares analysis of the heterogeneous effects of CAL treatment on student academic outcomes by student initial achievement level

Dependent variable: standardized evaluation test score (standard deviations)

	Math score		Chinese score	
	Fixed effect (1)	Random effects (2)	Fixed effect (3)	Random effects (4)
[1] CAL treatment in math only * Bottom 50% student in math (1=yes; 0=no)	0.01[c]	0.01[c]		
	(0.04)	(0.04)		
[2] CAL treatment in Chinese only * Bottom 50% student in Chinese (1=yes; 0=no)			0.09[d]	0.08[d]
			(0.06)	(0.06)
[3] Bottom 50% student (1=yes; 0=no) [a]	0.04	0.04	0.03	0.03
	(0.03)	(0.03)	(0.03)	(0.03)
[4] CAL treatment in math only	0.11**	0.11***	0.04	0.04
	(0.04)	(0.04)	(0.04)	(0.04)
[5] CAL treatment in Chinese only	0.25***	0.25***	0.13**	0.13**
	(0.06)	(0.06)	(0.06)	(0.06)
[6] Controls [b]	Yes	Yes	Yes	Yes
[7] Observations	15,421	15,421	15,421	15,421
[8] R-squared	0.330	0.327	0.347	0.339

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in brackets clustered at the school level.

Note: The test aims to show the heterogeneous effects of the different CAL treatments by student initial achievement level.

Columns (1) and (2) regress student standardized evaluation math test score on the main components and the interaction term of bottom 50% student in math (1=yes; 0=no) and indicator of CAL treatment in math only (1=only math treatment; 0=otherwise), and the main components and the interaction of the CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise) and bottom 50% student in math (1=yes; 0=no). Columns (3) and (4) regress students standardized evaluation Chinese test score on the main components and the interaction term of bottom 50% student in Chinese (1=yes; 0=no) and indicator of CAL treatment in Chinese only (1=only math treatment; 0=otherwise), and the main components and the interaction of the CAL treatment in both math and Chinese (1= both math and Chinese treatment; 0=otherwise) and bottom 50% student in Chinese (1=yes; 0=no). Columns (1) and (3) use the fixed effect model and columns (2) and (4) use the random effects model. All tests control for standardized baseline test score. All tests control for student characteristics that are listed in Appendix 1, rows (4)-(12).

[a] Bottom 50% student vary with the outcome variables. If the dependent variable is standardized evaluation Chinese test scores, then we use the indicator of bottom 50% student in Chinese. If the dependent variable is standardized evaluation math test scores, then we use the indicator of bottom 50% student in math.

[b] Control variables include all variables in rows (4)-(12) in Appendix 1. The baseline test scores we control for vary with the outcome variables. If the dependent variable is standardized evaluation math test scores, then we control for standardized baseline math test score. If the dependent variable is standardized evaluation Chinese test scores, then we control for standardized baseline Chinese test score. Also, indicators for the interaction of CAL treatment in both math and Chinese and student initial academic achievement served as controls in this analysis.

[c] To reach a significance level of 0.1, the two heterogeneous tests of the CAL treatment in math need to have a p-value of 0.05 *each* (using the Bonferroni method). In other words, the interaction term in row (1), columns (1)-(2), need to be significant at the 5% level after adjusting for multiple tests of heterogeneous effects. The results suggest that we cannot reject the hypothesis that CAL treatment in math only does not have heterogeneous effects by student initial achievement.

[d] To reach a significance level of 0.1, the two heterogeneous tests of the CAL treatment in Chinese need to have a p-value of 0.05 *each* (using the Bonferroni method). In other words, the interaction term in row (2), columns (3)-(4), need to be significant at the 5% level after adjusting for multiple tests of heterogeneous effects. The results suggest that we cannot reject the hypothesis that CAL treatment in Chinese only does not have heterogeneous effects by student initial achievement.

Appendix 1. Descriptive statistics of baseline characteristics of the treatment group and the control group of the sample students after attrition

		Students after attrition			
		Treatment group		Control group	
		Mean	Standard deviation	Mean	Standard deviation
[1]	Standardized baseline math test score (standard deviations)	0.02	1.00	0.02	0.98
[2]	Standardized baseline Chinese test score (standard deviations)	0.03	0.96	0.02	0.99
[3]	Standardized baseline total test score (standard deviations)	0.03	0.97	0.02	0.98
[4]	Gender (1=male; 0=female)	0.52	0.50	0.53	0.50
[5]	Only child (1=yes; 0=no)	0.23	0.42	0.23	0.42
[6]	Ever used a computer (1=yes; 0=no)	0.68	0.47	0.66	0.47
[7]	Father is illiterate (1=yes; 0=no)	0.09	0.29	0.10	0.29
[8]	Mother is illiterate (1=yes; 0=no)	0.23	0.42	0.24	0.43
[9]	At least one parent has an off-farm job (1=yes; 0=no)	0.35	0.48	0.37	0.48
[10]	Ever used internet (1=yes; 0=no)	0.29	0.45	0.28	0.45
[11]	Like school (1-100 points)	90.13	19.02	90.55	18.84
[12]	Baseline student self-efficacy (standard deviations)	-0.00	0.97	0.01	0.99
[13]	Observations	6919		8502	