Original article

doi: 10.1111/jcal.12106

Journal of Computer Assisted Learning

Persistence of learning gains from computer assisted learning: Experimental evidence from China

D. Mo,*† L. Zhang,‡ J. Wang,§ W. Huang,* Y. Shi,¶ M. Boswell* & S. Rozelle*

*Stanford University, Stanford, California, USA
†LICOS, University of Leuven, Leuven, Belgium
‡CCAP, IGSNRR, Chinese Academy of Sciences, Beijing, China
§Duke University, Durham, North Carolina, USA
¶Center for Experimental Economics in Education, Shaanxi Normal University, Xi'an, China

Abstract Computer assisted learning (CAL) programs have been shown to be effective in improving educational outcomes. However, the existing studies on CAL have almost all been conducted over a short period of time. There is very little evidence on how the impact evolves over time. In response, we conducted a clustered randomized experiment involving 2741 boarding students in 72 rural schools in China to evaluate impacts of CAL programs over the long term. Our results indicate that a CAL program that was implemented for one year and a half increased math scores by 0.25 standard deviations for third graders and 0.26 standard deviations for fifth graders. In addition, we have shown that students gained in math learning in both CAL Phase I (which ran for one semester in spring 2011) and CAL Phase II (which ran for both semesters of the 2011–2012 academic year) programs. By testing for heterogeneous effects, we find that the CAL intervention worked well for both the poorer performing and better performing students in the third and fifth grades. We also find that the third grade girls seem to have improved more than the boys in math in the short term (CAL Phase I).

Keywords Computer assisted learning, persistence of learning gains, randomized experiment, rural China.

Introduction

In the last decade, economists and education experts have studied the impact of computer assisted learning (CAL) programs on the educational performance of students in an attempt to help disadvantaged children in developing countries (e.g., Banerjee, Cole, Duflo, & Linden, 2007). These CAL programs utilize modern computing technologies to enhance learning through computerized instruction, drills and exercises in an environment where teaching and/or tutoring resources are in severe shortage (Barrow, Markman, & Rouse, 2009; Cuban & Kirkpatrick, 1998; He, Linden, & MacLeod, 2008; Linden, 2008). The software delivers simple teaching functions that might ordinarily be performed by a teacher (Stahl, Koschmann, & Suthers, 2006). For example, the software provides explanations about a student's curriculum and gives instructions about solving problems through animated lessons. The software can also provide feedback to students by correcting their answers to the exercises

Accepted: 11 January 2015
Correspondence: Di Mo, Freeman Spogli Institute, Stanford University, Encina Hall, 616 Serra Street, Stanford, CA 94305, USA. Email:dimo@stanford.edu

and illustrating the different approaches to getting the right answers. There have been several evaluations of CAL programs in developing countries that show positive impacts on students' performance (Banerjee *et al.*, 2007; Lai, Luo, Zhang, Huang, & Rozelle, 2011; Ong & Lai, 2006; Tüzün, Yılmaz-Soylu, Karakuş, İnal, & Kızılkaya, 2009).

However, an important limitation shared by nearly all such studies is that they were implemented over fairly short periods of time and therefore do not evaluate whether program impacts persist over time. For example, a study by Lai *et al.* (2011) only spanned a single semester, or 4 months. Similarly, studies by Banerjee *et al.* (2007) and Lai *et al.* (2012a, 2012b) encompassed only 9 months. In none of these studies did the research teams evaluate whether the program effect mainly took place in the first semester or whether it accumulated over the academic year. While these studies are helpful in exploring the impacts of a CAL program in the short run, they leave open interesting questions about the nature of the effect. Is the impact of a CAL program a 'one time' effect that diminishes once the novelty of computing wears off? Or, can CAL be considered a way to enhance learning that will continue to benefit students over the longer run?

The evidence on the question of whether CAL has persistent impact in the long run is mixed. Some previous studies point out that integrating regular class materials with interactive interfaces and computer-based games may make the learning process more engaging for students (Inal & Cagiltay, 2007; Schaefer & Warren, 2004). Such game-based learning software may increase students' motivation and interest in curricula, which may in turn lead to elevated focus and motivation among students (Chang, 2002; Cotton, 2002; Garris, Ahlers, & Driskell, 2002; Rico García & Vinagre Arias, 2000). However, there is little consensus on which game features actually support learning. The literature is also silent on the process by which games engage learners and the types of learning outcomes that can be achieved through gameplay (Garris *et al.*, 2002). Some studies raise the possibility that short-term gains in learning can be derived from the initial excitement of using a novel technology, but that this short-term gain may not be sustained over the long run (Lai *et al.*, 2011; Malone & Lepper, 1987; Marjanovic, 1999). What is more, after being exposed to a game-based curriculum students may find regular class periods boring and begin to disengage from the instruction of teachers (Cordova & Lepper, 1996). Under such circumstances, a long-term CAL program may appear to have no impact as reduced learning in regular classes might offset the gains achieved through CAL.

Unfortunately, there is little empirical evidence that can help us understand whether CAL programs can create significant and sustained gains in student learning in developing countries (Banerjee *et al.*, 2007). Many of the studies on computer-based learning and teaching programs do not allow for establishing causality due to the absence of a comparable control group (Blok, Oostdam, Otter, & Overmaat, 2002). Many other studies do not have an adequate sample size and, thus, lack statistical power. In the cases in which authors have used valid program evaluation techniques to study the long-run effects of computer-based learning, almost all have been conducted in developed countries and have targeted specific populations. For instance, Günther, Schäfer, Holzner, and Kemmler (2003) found that a computer assisted training program persistently improved cognitive abilities of elderly individuals with age-related memory deficits in the USA. Also in the USA, Roesch *et al.* (2003) found a positive long-term impact of an interactive computer program for medical case studies on dermatology students in medical school. Sustained gains in reading were also found among low-ability readers who participated in a computer-based reading program in France (Ecalle, Magnan, & Calmus, 2009).

These studies, notwithstanding important questions about the persistence of the impact of CAL programs on student learning in developing countries, remain unanswered. Because the quality of teacher resources is relatively poor in developing countries, and the demands on student learning are potentially greater due to rising student numbers in many parts of the developing world (Glewwe & Kremer, 2006), the absence of evidence on the long-term effects of CAL on students in developing countries warrants our attention. Were CAL programs to be integrated on a larger scale in developing countries, would the effects of CAL be positive and would they persist over time?

This question is particularly relevant to China. In order to narrow the 'digital divide' and the educational performance gap between rural and urban schools, China's Ministry of Education has an ambitious plan to invest in the computing infrastructure of rural schools (Ministry of Education, 2012). The recently announced 12th Five-Year Plan for Integrating Information Technology into Education aspires to set up a computer room in every rural school by 2020. Since the plan requires an enormous investment of fiscal resources, it is important to learn whether these resources can actually be made to promote sustained learning among rural students.

Central to the issue of learning sustainability and CAL programs is whether the impacts achieved through CAL derive from a particular piece of software or whether it is the 'act' of using the software that leads to change. Squires and McDougall (1994) adopted an evaluation framework seeking to measure the impact of software and computer packages on the educational performance of students. Other frameworks have been developed to understand how computers mediate learning in general (Tondeur, Van Braak, & Valcke, 2007). Among these, 'activity theory' suggests that Information and Communication Technology (ICT) may trigger changes in how subjects interact with one another as they learn (Lim, 2002). The theory posits that ICT tools, such as a desktop computer, help integrate (a) the subject – the individual student in our study; and (b) the object – higher math skills, by changing the way students approach tasks and assess their own learning progress, while also changing the format and the contents of leaning tasks. Lim (2002) also suggests that the activity system needs to be situated in a broader context to consider, for example, the school or local learning environment.

In this paper, we seek to use activity theory to assess if it is possible to effectively use game-based learning software to complement traditional ways of teaching in schools. There are several reasons to adopt this analytical approach. First, the program provides remedial tutoring in an environment where tutoring resources are absent (Lai *et al.*, 2012a). In rural schools in China, teachers often do not have time or energy to provide remedial tutoring. Parents commonly lack the adequate education to help their children. Commercial tutoring services rarely exist in rural areas. In such circumstances, computers may be an effective provider of remedial tutoring for rural students. In our intervention, we use CAL to provide remedial tutoring for learning gains. Second, when animations are used to present knowledge in computer software, learning interest and motivation may increase for young children (Szabo & Poohkay, 1996). The interest in learning may lead to long-term gains in learning. Third, computer shifts the learning activities from a traditional model of instruction to a learner-centred model that emphasizes a more active role for the learner (Garris *et al.*, 2002). New interactive technologies provide opportunities to create learning environments that more actively involve students in problem solving (Gokhale, 1996). The capabilities of active learning and problem solving may bring long-term benefits.

The overall goal of this study is to determine the persistence of CAL program effects on the academic outcomes of an underserved student population in a developing country. To achieve this goal, we pursue three specific objectives. First, we estimate whether a 1.5 year long math-based CAL program has any impact on academic performance. Second, we compare the program impacts after 1.5 years with those of a short-term program that ran for only one semester. Third, we explore the heterogeneous effects of the CAL intervention by investigating whether the treatment effects differ for different subgroups of students in both the short and the longer terms.

In order to achieve our objectives, we conducted the largest and longest lasting field experiment of CAL in China. The field experiment involved 72 rural schools and 2741 rural students. During the 1.5 year long experiment, we conducted a baseline survey and two rounds of evaluation surveys. The first evaluation was implemented one semester after the program started. The second evaluation was implemented three semesters, or 1.5 years, after the program started. Such a field experiment enables us to examine whether program impacts on student learning are persistent over time.

CAL program phase I – the short-term CAL program

In our first attempt to implement the CAL program as a short-term activity, we conducted a clustered randomized controlled trial (RCT) in Shaanxi Province in China in the spring of 2011. RCT is a type of experiment that is often used to test the efficacy and effectiveness of an intervention on a target population (Angrist & Pischke, 2014; Duflo, Glennerster, & Kremer, 2007). In such an experiment, subjects in the target population are randomly allocated to receive treatment or serve as a control, where they receive no treatment. This method minimizes allocation bias in testing the impact of the treatment on the target population.

A total of 5943 students in 72 rural schools in Shaanxi Province were involved in the study. Among these, 36 were intervention schools and 36 were control schools. Among the students, 2741 were boarders and the remaining 3202 were non-boarders. Boarding students are those students who live in a school-run dormitory between Monday and Friday of each school week. Boarding is optional and occurs at the discretion of parents, but is often necessary because the student's home is so far away from the school that commuting is infeasible. If a student does not board, he/she lives at home with his/her family and is called a non-boarder in this study. During the short-term CAL program, only the boarding students in the 36 treatment schools participated in the CAL classes, while the non-boarders served as additional controls.

The short-term CAL intervention that was implemented in the 36 treatment schools ran for one 4-month semester. In basic terms, the CAL program consisted of a remedial, game-based CAL program in math that was held outside of regular school hours among boarding students. Complete details of the CAL intervention are included in the next section. To test the effectiveness of the CAL program, students in both treatment and control schools were given standardized math tests before the start of the program and at the end of its implementation. The standardized math test included questions from math exercise books that are available in bookstores. The questions were chosen by educational testing specialists for primary education in China. The questions were chosen to test the math knowledge and skills that students should master according to the national curriculum. The tests were administered in the same manner to all sample students in both treatment and control groups. Different tests were given to students in different grades.

According to the analysis presented in Figure 1 and Table 1 (and also in Lai *et al.*, 2012a), the short-term CAL program had a positive and significant effect on the math test scores of students in the treatment schools. Overall, scores went up by 0.12 standard deviation (Table 1). Table 1 includes the results of the regressions when using the full model and the exact specifications can be found in the section below. The results show that the CAL treatment effect of phase I on all students, including both third and fifth grade students, is 0.12 standard deviation (Table 1, row 1, column 1). The estimated coefficient is significant at the 5% level. The results also indicate that the treatment group improved more than the control group by 0.12 standard deviation in math score due to the treatment. By dividing the sample into the third and the fifth grades, it is shown that the impact on the third grade students is 0.18 standard deviation (row 1, column 2). The estimated coefficient is significant at the 5% level. The estimated coefficient of the treatment effect on the fifth grade students only is 0.07 standard deviation and is not significant at the 10% level (row 1, column 3).

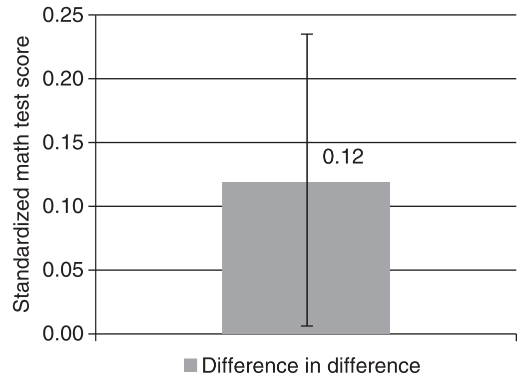


Figure 1 Difference in Difference in the Standardized Math Test Scores before and after the CAL Phase I Program (March 2011 and June 2011) between the Treatment and Control Groups in Both the Third and the Fifth Grades (cited from Lai *et al.*, 2012a)
Note: Difference in difference methodology measures the differences in the improvement in math test scores during the program period between the treatment and control groups before and after the treatment. In carrying out the difference approach, we first calculated the difference of the math test scores of every student (both treatment and control) between the pre- and post-program periods. We then calculate the difference in the average improvement of the treatment and control group students. When executing the difference in difference methodology in this way, the bar graph illustrates the impact of the CAL treatment on the improvement in student math scores during the program period.

Despite the positive result in Figure 1 and Table 1, there remains the question of whether the impact of the phase I short-term program would persist over a longer period of time. If the entire school system were to adopt this program, it would be essential to first learn whether the observed findings represent only a short-term impact or whether the program effect can in fact

Table 1. Ordinary Least Square Analysis of CAL Phase I (March–June 2011) on Students' Standardized Math Test Scores for Third and Fifth Grade Students

		All (third grade and fifth grade) (1)	Third grade (2)	Fifth grade (3)
[1]	Treatment (1 = treatment group; 0 = control group)	0.12**	0.18**	0.07
		[0.05]	[0.08]	[0.07]
[3]	Control variables	Yes	Yes	Yes
[4]	Observations	2613	1124	1489
[5]	R^2	0.26	0.29	0.25

Note. Robust standard errors in brackets clustered at school level. Cited from Lai *et al.* (2012a).
**Significant at 5%.

be sustained. That is why we designed the longer term second phase of the experiment.

Sampling, data and methods for CAL phase II – the longer term CAL program

Sampling and the process of randomization

For phase II of the experiment, we conducted a school level clustered RCT of the CAL program in Shaanxi rural schools during the entire 2011–2012 academic school year. Each academic year is divided into the spring semester and the fall semester with 4 months per semester. The second phase of the CAL program was implemented as an extension of the first phase, which ran for the duration of the spring semester of 2011 (as discussed above). As such, CAL phase II included the same 2741 boarding students in the same 72 rural schools in Shaanxi Province. The sample students were in third and fifth grades when they joined phase I in the spring of 2011. During phase II (October 2011–June 2012), the students were in fourth and sixth grades.

Choosing the sample consisted of several steps. First, to focus our study on poor rural students, we restricted our sample frame to four counties randomly selected out of the ten counties in Ankang Prefecture, an administrative area that covers a poor region in southern Shaanxi Province. Shaanxi Province has 40 million people, about 60 million of whom live in rural areas. In 2011, the average annual per capita income in the sample counties was approximately 4000 RMB ($US600), compared with rural China's average per capita income of 6977 RMB in the same year [China National Bureau of Statistics (CNBS), 2011]. After selecting the counties, we obtained a comprehensive list of all elementary schools with grades 1–6 in the four counties. We selected all 72 schools that met these criteria to be our sample.

During phase I, the sample included both third grade and fifth grade students. We chose students from these grades for several reasons. First, at the time of the launch of the project in spring 2011, we only had remedial tutoring materials for grades 3–6 and thus did not choose students from the first or second grade. Second, given the limited number of computers in each school's computer room and the scheduling constraints of boarding students, the CAL program could only accommodate students from two grade levels. We excluded the sixth grade students from consideration because they would have graduated before the phase II program had begun. Being two grades apart, the third and fifth graders could also offer a sharper comparison of the intervention effects by age group. None of these students had ever participated in a CAL program prior to the spring semester of 2011.

All boarding students in the 72 sampled schools were included in the sample. In the spring semester of 2011, there were a total of 2741 boarding students in the sample, of which 1167 were in the third grade and 1574 were in the fifth grade (Figure 2). We chose boarding students for the sample because they are among the most vulnerable students in rural China (Lai *et al.*, 2012a). Our sample selection criterion is consistent with the general goal of targeting the vulnerable student populations in rural China. Our previous studies have shown that rural boarding students perform significantly worse academically than rural

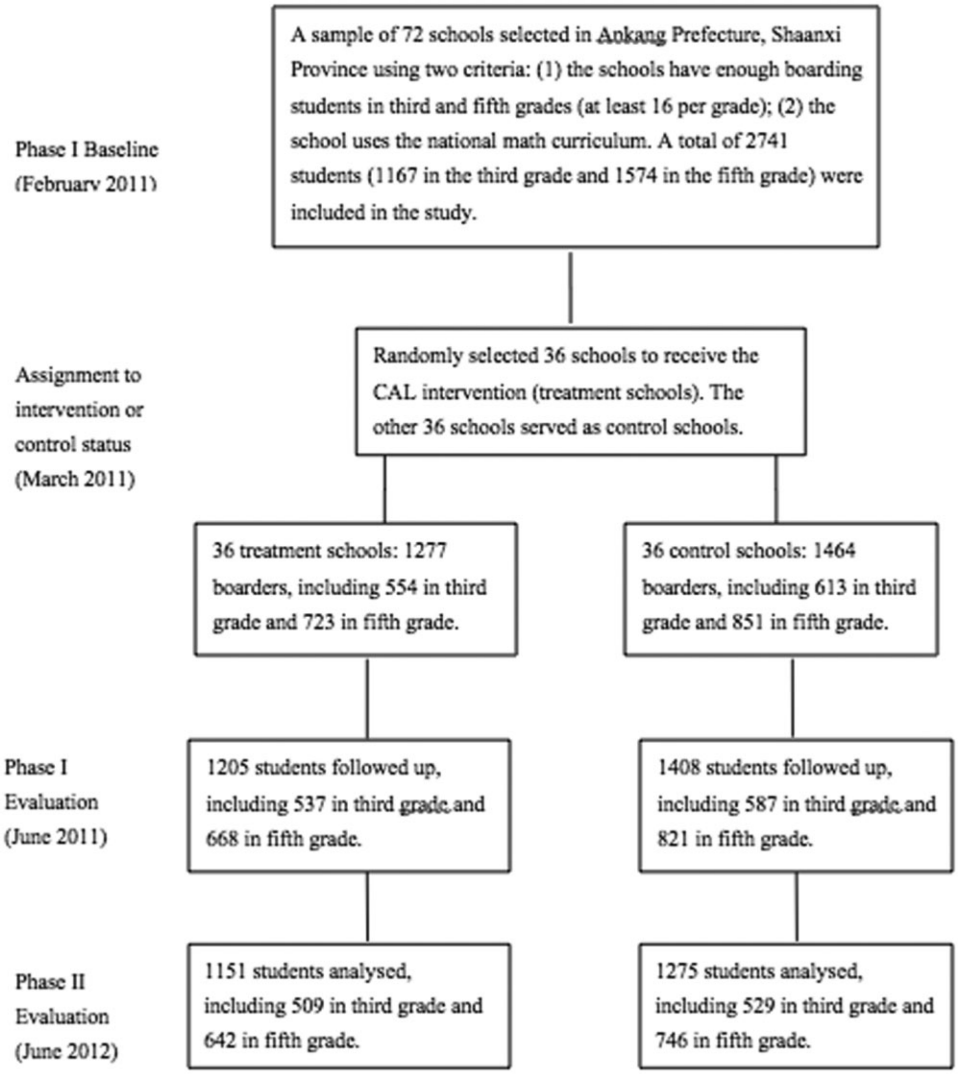


Figure 2 Experiment Profile

non-boarding students and are more likely to suffer from health and nutrition problems (Luo *et al.*, 2012; Mo *et al.*, 2012).

Although there was some sample attrition by the end of phase II, it is unlikely to have had a large impact on the study. Because of school transfers, illness or injury, 11.5% of the students in the sample attritted between the baseline and endline surveys. This attrition rate is low compared with other experiments conducted among primary schools in developing countries (McEwan, 2013). By the time of the final evaluation of the phase II program, survey enumerators were able to follow up with 2426 boarding students in the 72 sample schools (Figure 2, final row). There were 129 attrited students (11.1%) from the third grade and 186 attrited students (11.8%) from the fifth grade.

Fortunately for the study's integrity, there was almost no systematic relationship between the treatment and attrition status or student characteristics and attrition status (Table 2). In other words, among the attrited students, there were no student-level variables that were correlated in a statistically significant way with the treatment/control status of the students (column 1). The treatment students were as likely as the control students to attrit, and attrited students had similar characteristics in both groups.

We randomly chose 36 schools from the 72 schools in our sample to receive the CAL intervention. All of

Table 2. Comparison of Students' Characteristics between Attrited and Non-Attrited Students in the Sample and between Treatment and Control Students during the Entire CAL Intervention (Phase I and Phase II)

		Difference between attrited and non-attrited students[c]		Difference between the treatment and control groups within attrited students
		Third grade (1)	Fifth grade (2)	All attrited students[d] (3)
[1]	Baseline math score (units of standard deviation)[a]	–0.01	–0.03	0.02
		(0.01)	(0.02)	(0.04)
[2]	Baseline Chinese score (units of standard deviation)[b]	–0.02**	–0.01	–0.05
		(0.01)	(0.01)	(0.04)
[3]	Gender (1 = boy; 0 = girl)	–0.01	0.03*	0.03
		(0.02)	(0.02)	(0.05)
[4]	Age (years)	0.01	–0.00	0.01
		(0.01)	(0.01)	(0.03)
[5]	Only child (1 = yes; 0 = no)	0.01	0.03**	0.02
		(0.02)	(0.02)	(0.05)
[6]	Ever repeated grade (1 = yes; 0 = no)	0.02	0.02	–0.07
		(0.02)	(0.02)	(0.06)
[7]	At least one parent has junior high school or higher degrees (1 = yes; 0 = no)	0.02	0.01	0.06
		(0.02)	(0.02)	(0.06)
[8]	At least one parent has senior high school or higher degrees (1 = yes; 0 = no)	0.03	–0.01	–0.02
		(0.03)	(0.03)	(0.09)
[9]	At least one parent has an off-farm job (1 = yes; 0 = no)	0.02	–0.04*	0.03
		(0.02)	(0.02)	(0.08)
[10]	Family wealth (1 = higher than the median; 0 = otherwise)	0.02	0.02	–0.08
		(0.02)	(0.02)	(0.06)
[11]	Ever used a computer (1 = yes; 0 = no)	–0.05	–0.05	0.15
		(0.04)	(0.08)	(0.12)
[12]	Observations	1167	1574	305

[ab]The baseline math score is the score on the standardized math test that is given to all sample students before the CAL program. [c]The sample includes both the sample observations (non-attrition) and the attrition observations. [d]The sample is limited to the attrited observations.
*Significant at 10%; **Significant at 5%. Robust standard errors in brackets clustered at school level.

Table 3. Difference in Characteristics between Students in the Treatment Group and the Control Group during the Entire CAL Intervention (Phase I and Phase II)

Dependent variable: whether the student received CAL treatment (1 = yes; 0 = no)

		Third grade[c] (1)	Fifth grade[c] (2)
[1]	Baseline math score (units of standard deviation)[a]	0.02	–0.04
		(0.03)	(0.04)
[2]	Baseline Chinese score (units of standard deviation)[b]	–0.04	0.01
		(0.03)	(0.03)
[3]	Gender (1 = boy; 0 = girl)	–0.02	–0.03
		(0.03)	(0.03)
[4]	Age (years)	0.00	0.04
		(0.03)	(0.02)
[5]	Only child (1 = yes; 0 = no)	0.02	–0.05
		(0.05)	(0.03)
[6]	Ever repeated grade (1 = yes; 0 = no)	0.00	–0.00
		(0.04)	(0.04)
[7]	At least one parent has junior high school or higher degrees (1 = yes; 0 = no)	0.01	–0.03
		(0.03)	(0.03)
[8]	At least one parent has senior high school or higher degrees (1 = yes; 0 = no)	0.06	0.04
		(0.05)	(0.05)
[9]	At least one parent has an off-farm job (1 = yes; 0 = no)	–0.05	–0.02
		(0.05)	(0.04)
[10]	Family wealth (1 = higher than the median; 0 = otherwise)	–0.01	–0.00
		(0.04)	(0.04)
[11]	Ever used a computer (1 = yes; 0 = no)	0.05	0.08
		(0.11)	(0.12)
[12]	Observations	1,038	1,388
[13]	R^2	0.111	0.120

Note. Robust standard errors in brackets clustered at school level.
[ab]The baseline math score is the score on the standardized math test that is given to all sample students before the CAL program.
[c]The sample includes the remaining sample (non-attrition).

the 1277 boarding students in the third and fifth grades of the 36 treatment schools constituted the treatment group (Figure 2). Among these students, there were 554 third grade students and 723 fifth grade students. The 1464 boarding students in the same grades in the other 36 schools, including 613 from the third grade and 851 from the fifth grade, served as the control group. Because of attrition, there were 2426 students left in the final analytic sample, among whom 1151 were in the 36 treatment schools and 1275 were in the control schools (Figure 2).

The balance of the sample across treatment and control groups was also even (Table 3). To show this, a set of student characteristics can be used to check the validity of the random assignment. As is standard in the program evaluation literature, the treatment variable is regressed (*whether the student received CAL treatment or not*) on the characteristics of the students. According to the data, none of the differences in student characteristics between the treatment and control groups were statistically significant (columns 1 and 2). In addition, almost all the differences between treatment groups are small in magnitude.

Intervention

The main intervention involved computer assisted math remedial tutoring sessions, which were designed to complement the regular in-class math curriculum for the spring semester of 2011 (phase I) and the entire school year of 2011–2012 (phase II). During this program, the CAL sessions were given to the students under the monitoring of two teacher-supervisors trained by our research group. The students in the treatment group participated in two 40-min CAL sessions per week. The sessions were mandatory and attendance was taken by the teacher-supervisors.

The content of each session included instructional videos and games and was designed to help the students reach basic competencies in China's uniform national math curriculum. The software-based lessons were exactly the same for all students of the same grade in each of the treatment schools.

During each CAL session, students sat at computers and played math games designed to help them review and practise the basic math material that was being taught in their regular school math classes. The CAL teacher-supervisors arranged for the students to sit in pairs, with one pair of students sharing a single computer. The students shared one pair of earbuds so that each could hear the voices, music and other sounds of the software. Only one student at any given time had control of the mouse, but at regular intervals the students were encouraged to take turns using the mouse.

In a typical session, the students first watched an animated video that reviewed the material on which they were receiving instruction during their regular math class sessions in that same week. The students then played math games with animated characters to practise the skills introduced in the video lecture. If a student had a math-related question, he/she was encouraged to discuss the question with his/her teammate with whom he/she shared the computer. The students were not allowed to discuss with other teams or the teacher-supervisor. Generally, the games involved an animated character engaged in some task, such as archery and crossing a river. Multiple-choice questions would then appear on the screen one at a time. Successful answers would aid the animated character in their task and incorrect answers would trip them up and/or slow them down. Either way, humorous animations would appear once the students chose their response. Both students in each pair had access to scratch paper at their station to take notes and make calculations. At the end of each game, students were shown how many of the questions had been answered correctly.

Our protocol required that the teachers could only help students with scheduling, computer hardware issues and software operations. This was done to try to control for the possible effect of the CAL supervisor's involvement and thereby to make sure any observed impact would be entirely due to the CAL program itself. In fact, according to our observations during occasional unannounced visits to randomly selected schools, the software demanded the full attention of students. There was little, if any, interaction between the students and the teacher-supervisors. In addition, while there was a lot of interaction within each of the two-person teams, there was little communication between pair groups.

The intervention team spent considerable time in preparing the necessary hardware, software, curriculum and program implementation protocol in a way that would both facilitate smooth implementation of the CAL program and avoid confounding influences that might bias our results. As the first step, to meet the hardware requirements of the CAL program, Dell Inc. donated 640 brand new identical desktop computers and research staff installed the CAL software package on these desktops. These staff also removed all pre-installed software that would not be used during the CAL intervention (such as Windows built-in games and Microsoft Office) and disabled the Internet and USB functions on all of the computers. Setting up computers in this way prevented students and teachers from using the program computers for other purposes that might affect the operation of the regular CAL program, as well as the interruptions that might be caused by accidental deletion of the CAL software or the introduction of viruses. Sealing the computers also ensured the quality of our evaluation of the program effects without capturing any other confounding influences, such as students looking up answers on the Internet. It also prevented teachers/students in control groups from copying our CAL software onto other computers.

All teacher-supervisors of the 36 treatment schools also participated in a 2-day mandatory training program. The training was designed to prepare the teacher-supervisors for their responsibilities in the CAL classes. The teacher-supervisors' five main responsibilities included: (a) taking attendance; (b) making sure that the CAL curriculum in each session was matched to the curriculum being taught in the students' math class; (c) managing the CAL classrooms so that order was maintained; (d) providing immediate assistance when students experienced difficulty in computer and/or math game software operations (but they were not to instruct the students in math); and (e) taking care of the CAL desktops and keeping close contact with the research group/volunteers regarding technical support or CAL management questions. Because this work was clearly beyond the scope of their normal classroom duties, we compensated the teacher-supervisors with a monthly stipend of 100 yuan (approximately $US15). This is an amount roughly equivalent to 15% of the wage of a typical rural teacher.

CAL control group
The third and fifth grade boarding students in the 36 control schools constituted the CAL control group. Students in the control group did not receive any CAL intervention. To avoid spillover effects from the CAL intervention, the principals, teachers and students of the control schools were not informed of the CAL project. The research team did not visit the control schools except during the baseline and final evaluation surveys. Informed assent was given by the guardian of the students so that the children could participate in the baseline and endline surveys. However, neither students nor teachers in the control group knew that there were students in other schools participating in the CAL program. The students in the control group took their regular math classes at school as usual.

Data collection

The research group conducted three rounds of surveys in the 72 control and treatment schools. The first-round survey was a baseline survey conducted with all third and fifth grade boarding students in the 72 schools in late February 2011 at the beginning of the spring semester and before any implementation of CAL program had begun. The second-round survey was an evaluation survey conducted in June 2011, a time that coincided with the end of the spring semester of 2011 and the end of phase I. The third-round survey was a final evaluation conducted at the end of phase II in June 2012.

In each round of the survey, the enumeration team visited all schools and conducted a two-part data collection effort. In carrying out this effort, the team gave a math test to the students as well as a survey. The main outcome variable in the study is derived from the in-class math test. The test was the same for all students in the same grade across groups and schools in each round of survey. Students were required to finish the test in 25 min. Although drawn out of the same pool of questions, the math questions were different for each round of the survey. Math questions in the test did not repeat the exercises included in the CAL software. Our enumeration team closely monitored the test and strictly enforced the time limits. In order to make test scores from different rounds of survey comparable, scores are normalized relative to the distribution of the baseline test scores of the control group. Specifically, we subtracted the mean of the control group in the baseline and divided by the standard deviation of the control group in the baseline.

After the math test, enumerators conducted a survey to collect data on the characteristics of students and their families. This part of the survey formed the basis for a set of demographic and socio-economic variables. The dataset includes measures of each student's *gender*, *age* (measured in years), *only child* (if the student is the only child of his or her family), grade repetition (if the student has *ever repeated a grade*), parents' education level (*at least one parent has junior high school or higher degree* and *at least one parent has senior high school or higher degree*), parents' job (*at least one parent has an off-farm job*), *family wealth* (the variable of family wealth equals 1 if the family assets are higher than the median value and 0 otherwise) and computer use (the variable equals 1 if the student had *ever used a computer*).

The questionnaires used in the survey were designed to collect baseline values of covariates that are likely to influence or predict the outcome – student's math test score. It has been shown that controlling for baseline covariates that have large effects on the outcomes in a study's regression analysis can help reduce standard errors of the estimates and increase power (Duflo *et al.*, 2007).

There are a number of reasons why we chose the set of covariates in our final specification. Both the theoretical and empirical literatures suggest that student's characteristics are good predictors of student's test scores. For example, a number of authors find that gender plays an important role in predicting math test scores (Niederle & Vesterlund, 2010). Age has been shown to be a significant variable in other studies (Salamonson & Andrew, 2006). Other research teams have found that family background variables are important indicators of educational success (Woessmann, 2004). Still in others, it has been found that family wealth and parental education can help predict school attainment in China (Zhao & Glewwe, 2010). Therefore, based on the existing both theoretical and empirical literatures, we surveyed the students and collected data from students in order to be able to create a set of demographic and socio-economic variables that may be good predictors of math test score and help improve the precisions of our impact evaluation results.

Statistical methods

Unadjusted and adjusted ordinary least squares regression analyses estimate how the academic performance changed in the treatment group relative to the control group. Unadjusted analysis regressed changes in the outcome variable (i.e., post-program math test score minus baseline math test score) on a dummy variable of the treatment (CAL intervention) status. Adjusted analyses improve statistical efficiency – these approaches are described in detail in the models below. The study randomizes treatment at the school level. Therefore, it is possible that the error terms are correlated within schools. To account for intra-cluster correlation, standard errors clustered at the school level are used in all the regressions testing the treatment effect. See Imbens and Wooldridge (2008) for more details.

The unadjusted model is:

$$\Delta y_{is} = \alpha + \beta \cdot treatment_s + e_{is} \quad (1)$$

where Δy_{is} is the change in the outcome variable during the program period for child i in school s, $treatment_s$ is a dummy variable for treatment school students (equal to 1 for students in the treatment group and 0 otherwise), and ϵ_{is} is a random disturbance term clustered at the school level. By construction, the coefficient of the dummy variable $treatment_s$, β, is equal to the unconditional difference in the change in the outcome (Δy_{is}) between the treatment and control groups over the program period. In other words, β measures how the treatment group changed in the outcome levels during the program period relative to the control group.

In order to improve the efficiency of the estimation, we built on the unadjusted model in equation (1) by including a set of control variables:

$$\Delta y_{is} = \alpha + \beta \cdot treatment_s + \theta \cdot y_{0is} + X_{is}\gamma + \varepsilon_{is} \quad (2)$$

where all the variables and parameters are the same as those in equation (1), except that we added a set of control variables. Specifically, we control for y_{0is}, the baseline math test scores for student i in school s, and X_{is}, a vector of additional control variables. The variables in X_{is} are student and family characteristics (gender, age, only child, ever repeated grade, at least one parent has junior high school or higher degrees, at least one parent has senior high school or higher degrees, at least one parent has an off-farm job, family wealth, ever used a computer). By including y_{0is} and X_{is} as control variables, β in equation (2) provides an unbiased, efficient estimate of the CAL treatment effect.

Estimates of the treatment effect of the CAL intervention occur across three time horizons. In order to show the longer term effects of the CAL program, the first estimation is of the longer term treatment effect of three semesters (CAL phase I and phase II). The second estimation is of the treatment effect for phase II only and can be compared with the effect of phase I. This approach shows how the CAL program effect evolves from the one-semester program (phase I) to the two-semester program (phase II). Both equations (1) and (2) were used in estimating treatment effects across the three time horizons.

Results

The data show that students in the treatment group improved significantly more in their math performance than did students in the control group after taking the CAL classes for three semesters, from the beginning of the spring semester 2011 to the end of the spring semester 2012. The students in the treatment group and the control group in the third grade started at similar levels in pre-test standardized math scores at the start of spring semester of 2011 (Figure 3, panel a). After three semesters of treatment, the treatment group improved more in math than did the control group (panel a). The difference in the change in standardized math test scores between the two groups was 0.21 standard deviation for the third grade students (panel b).

The results are similar for the fifth grade students (Figure 4). The data show that students in the fifth grade also improved significantly more in terms of their standardized math test scores than the students in the control group after three semesters of taking the CAL classes. In a statistical sense, the fifth grade baseline standardized test scores of control students are the

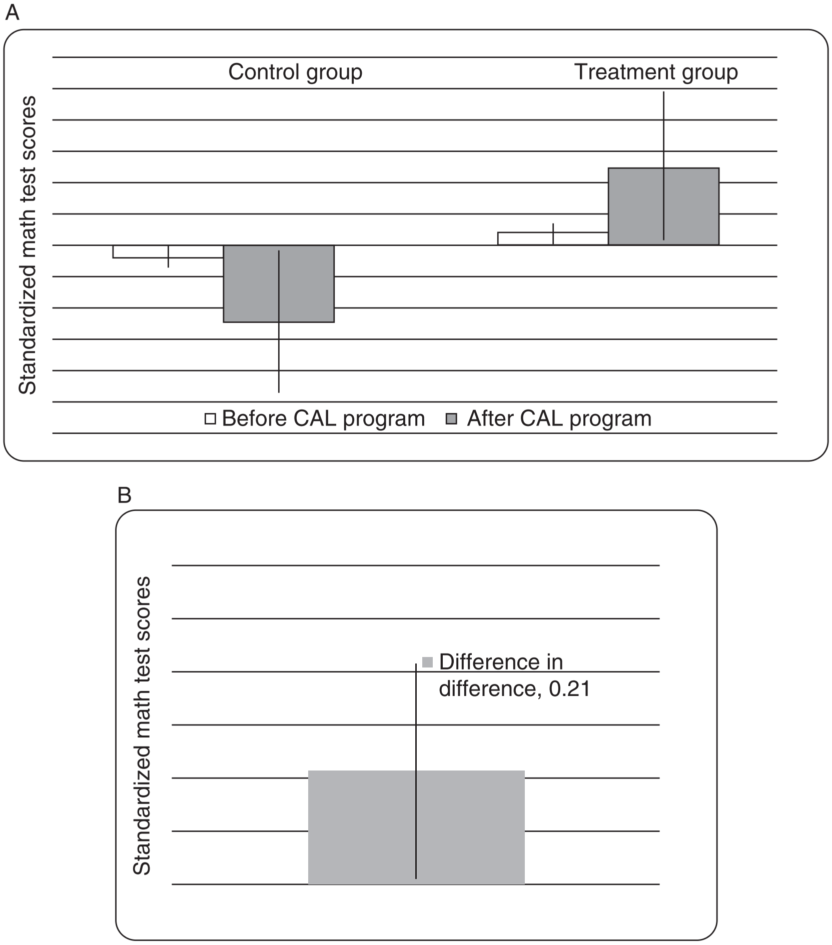


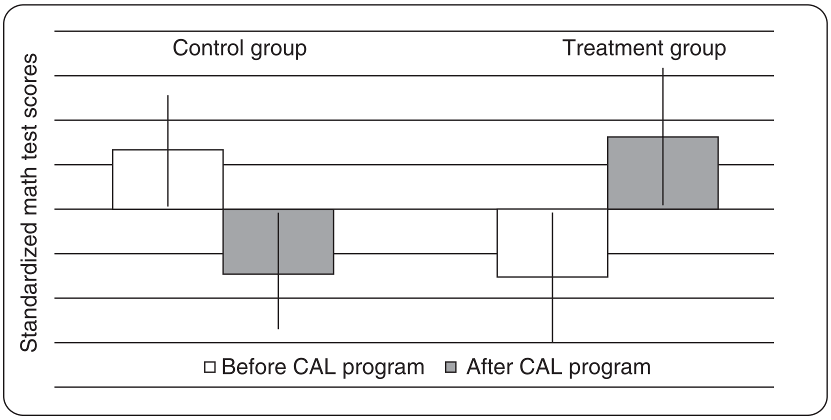
Figure 3 Change in the Standardized Math Test Scores of Third Grade Students before and after the Entire CAL Intervention (Phase I and Phase II). (A) Standardized Math Test Scores before and after the Entire CAL Intervention (Phase I and Phase II): Treatment and Control Groups in Third Grade. (B) Difference in Difference in the Standardized Math Test Scores before and after the Entire CAL Intervention (Phase I and Phase II) between Treatment and Control Groups in Third Grade

same as the scores of the treatment students (Figure 4, panel a). After the CAL intervention of three semesters, the students in the treatment group improved by 0.29 standard deviation more than did the students in the control group (panel b).

The adjusted and unadjusted multivariate regression analyses are consistent with our graphical descriptive analysis (Table 4). Using only the third grade students or only the fifth grade students, the estimated CAL treatment effects on math test scores using results from the unadjusted model are 0.21 standard deviation for the third grade students (row 1, column 1) and 0.30 standard deviation for the fifth grade students (row 1, column 2). The estimated treatment effects for both grades are statistically significant at the 10% level in the case of the third grade cohort and the 1% level in the case of the fifth grade cohort.

When we add the additional control variables, using the adjusted model, the results from the more efficient estimator demonstrate that the treatment effect is still large and statistically significant (Table 4). In the case of the third grade students, the estimated treatment effect is 0.25 standard deviation (row 1, column 3). In the case of the fifth grade students, the estimated treatment effect is 0.26 standard deviation. Both of the estimates are significant at the 1% level (row 1, columns 3 and 4). An increase of one-fourth of a standard deviation can amount to a considerable gain in performance. Such an increase in performance is estimated by some to be equivalent to 0.6 year of schooling (Glewwe, Park, & Zhao, 2011). A similar effect size was found in other prominent education experiments. For example, the Tennessee Star Program sought to measure the effect of reducing class size by one-third (from a classroom of 22–25 students to a classroom of 13–17 students; Mosteller, 1997). The program was considered successful in that test scores were raised by 0.25 standard deviation. The measured effect for the CAL program size is similar in magnitude.

A



B

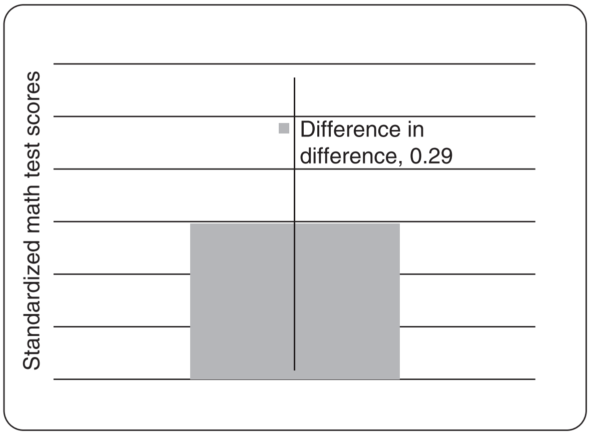


Figure 4 Change in the Standardized Math Test Scores of Fifth Grade Students before and after the Entire CAL Intervention (Phase I and Phase II). (A) Standardized Math Test Scores before and after the Entire CAL Intervention (Phase I and Phase II): Treatment and Control Groups in Fifth Grade. (B) Difference in Difference in the Standardized Math Test Scores before and after the Entire CAL Intervention (Phase I and Phase II) between Treatment and Control Groups in Fifth Grade

Comparing the effects of the CAL intervention between phase I and phase II

Using the results from the regression model based on the specification in equation (2), the results show that the effect of the CAL treatment appears to persist in the longer run (Table 5). Our point estimate shows that during CAL phase I (March 2011–June 2011), the estimated treatment effect for the third grade boarding students is equal to 0.18 standard deviation and is significant at the 5% level (row 1, column 1). During CAL phase II (September 2011–June 2012), the point estimate of the treatment effect for the third grade students is still positive. The magnitude is 0.07 standard deviation, although statistically insignificant (row 1, column 3). Hence, this result (0.25 standard deviation shift over the two study phases as shown in Table 4) suggests that the impact persisted in the longer term.

Consistent with the third grade, the fifth grade boarding students also improved in the longer term CAL program. The estimated treatment effect of phase I is equal to 0.11 standard deviation, although this is statistically insignificant (Table 5, row 1, column 2). However, during phase II the estimated treatment effect becomes significant at the 10% level and the point estimate is equal to 0.15 standard deviation (row 1, column 4). Like the third graders, fifth graders improved in both the short term and the longer term to achieve an overall learning improvement of 0.26 standard deviation after three semesters of CAL classes.

The high interest level in the CAL software among the treatment students supports these results. The ratings of student interest in the software (0–100 points) at the end of CAL phase II suggest that the students were highly interested in the software regardless of their previous computer experience or academic performance (Appendix I). The mean rating of student interest was 88 points for the third grade students and 83 points for the fifth grade students. Third grade students who had used computers before the CAL

Table 4. Ordinary Least Squares Estimators of the Impacts of the Entire CAL Intervention (Phase I and Phase II) on Students' Math Test Scores

Dependent variable: standardized post-CAL math test score – standardized baseline math test score (standard deviations)

		Third grade (1)	Fifth grade (2)	Third grade (3)	Fifth grade (4)
[1]	Treatment (1 = treatment group; 0 = control group)	0.21*	0.30***	0.25***	0.26***
		(0.11)	(0.10)	(0.08)	(0.08)
[2]	Baseline math score (units of standard deviation)[a]			–0.64***	–0.59***
				(0.04)	(0.03)
[3]	Baseline Chinese score (units of standard deviation)[b]			0.18***	0.15***
				(0.04)	(0.03)
[4]	Gender (1 = boy; 0 = girl)			0.13**	0.02
				(0.05)	(0.04)
[5]	Age (years)			–0.11**	–0.10***
				(0.04)	(0.03)
[6]	Only child (1 = yes; 0 = no)			0.06	0.02
				(0.08)	(0.05)
[7]	Ever repeated grade (1 = yes; 0 = no)			–0.03	0.04
				(0.06)	(0.05)
[8]	At least one parent has junior high school or higher degrees (1 = yes; 0 = no)			0.00	0.04
				(0.07)	(0.04)
[9]	At least one parent has senior high school or higher degrees (1 = yes; 0 = no)			–0.07	–0.05
				(0.08)	(0.08)
[10]	At least one parent has an off-farm job (1 = yes; 0 = no)			–0.06	0.04
				(0.07)	(0.07)
[11]	Family wealth (1 = higher than the median; 0 = otherwise)			–0.07	0.10**
				(0.05)	(0.04)
[12]	Ever used a computer (1 = yes; 0 = no)			0.07	0.25*
				(0.10)	(0.14)
[13]	Observations	1038	1388	1038	1388
[14]	R^2	0.011	0.024	0.322	0.293

Note. Robust standard errors in brackets clustered at school level.
[ab]The baseline math score is the score on the standardized math test that is given to all sample students before the CAL Program.
*Significant at 10%; **Significant at 5%; ***Significant at 1%.

Table 5. Ordinary Least Squares Estimators of the Impacts of the CAL Program during Phase I and Phase II on Students' Math Test Scores for Third and Fifth Graders

Dependent variable: standardized post-CAL math test score – standardized baseline math test score		CAL phase I		CAL phase II	
		Third grade (1)	Fifth grade (2)	Third grade (3)	Fifth grade (4)
[1]	Treatment (1 = treatment group; 0 = control group)	0.18**[b]	0.11[c]	0.07[b]	0.15*[c]
		(0.08)	(0.07)	(0.10)	(0.08)
[2]	Control variables[a]	Yes	Yes	Yes	Yes
[3]	Observations	1038	1388	1038	1388
[4]	R^2	0.301	0.261	0.048	0.038

Note. Robust standard errors in brackets clustered at school level.
[a]Control variables include all the variables in Table 2 and the township dummies. [b]Wald test shows that the treatment effect on the third grade students is not significantly different between phase I and phase II (p value = 0.48). [c]Wald test shows that the treatment effect on the fifth grade students is not significantly different between phase I and phase II (p value = 0.73).
*Significant at 10%; **Significant at 5%.

program had an interest rating as high as 88 points (row 1, column 2). The third grade students without any computer experience before CAL had a rating of 89 points (row 1, column 3). However, the difference is not significant (row 1, columns 4–5). The difference between the fifth grade students with and without previous computer experience is also small and not significant (row 2, columns 2–5). Moreover, both the better and worse performing third grade and fifth grade students showed high interest in the software, although none of the differences were significant (rows 1 and 2, columns 6–9). These findings suggest that all students, regardless of time exposed to computer technology and level of academic performance, maintained a high level of interest in CAL.

Heterogeneous effects of the CAL intervention

In order to test for heterogeneous program effects, included in the regression model specified in equation (2) are interaction terms between the treatment dummy variable and two key covariates. Including in the regression, an interaction term between the treatment dummy variable and the variable of *baseline math test score* reveals, for example, whether the change in math test scores differed for students who were better performing in math at the time of the baseline relative to students who were poorer performing. Similarly, including the interaction term between the treatment dummy variable and the variable of *gender* tests whether boys benefited differently from the program than girls.

The estimated results using equation (2) – which includes the interaction term between the treatment dummy variable and the baseline math test score – demonstrate that the CAL intervention worked similarly well for the better performing and poorer performing students in the third and fifth grades (Table 6). There is no significant evidence of heterogeneous program effects of CAL on standardized math test scores (row 2). Students in the third and fifth grades who scored relatively high and relatively low on the baseline math test did equally well after the entire CAL treatment of three semesters (row 2, columns 1 and 4). When estimating with phase I only and phase II only, the poorer and better performing students have no significant differences in their math improvements (row 2, columns 2, 3, 5 and 6).

However, there does seem to be an interesting heterogeneous effect among girls and boys in the third grade (Table 7). The third grade girls improved more than the boys by 0.21 standard deviation in math in phase I (the difference is significant at the 5% level, row 2, column 2). More specifically, after the CAL intervention in phase I, girls in the treatment group improved by 0.29 standard deviation in math relative to the girls in the control group (row 1, column 2), while the boys in the treatment group improved by 0.08 (0.29–0.21) standard deviation in math relative to the boys in the control group (rows 1 and 2, column 2). The difference in math improvement between girls and boys is ultimately reduced as boys seem to catch up by improving more than the girls in phase II where the

Table 6. Ordinary Least Squares Estimators of the Heterogeneous Effects of CAL Intervention on Standardized Math Test Scores by Baseline Math Performance

Dependent variable: standardized post-CAL math test score – standardized baseline math test score

		Third grade			Fifth grade		
		Phase I and Phase II (1)	Phase I (2)	Phase II (3)	Phase I and Phase II (4)	Phase I (5)	Phase II (6)
[1]	Treatment (1 = treatment group; 0 = control group)	0.25***	0.19**	0.07	0.26***	0.10	0.16**
		(0.08)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)
[2]	Treatment × Standardized baseline math test score	0.01	0.04	–0.03	–0.03	–0.07	0.05
		(0.06)	(0.06)	(0.09)	(0.06)	(0.06)	(0.07)
[3]	Control variables[a]	Yes	Yes	Yes	Yes	Yes	Yes
[4]	Observations	1038	1038	1038	1388	1388	1388
[5]	R^2	0.322	0.301	0.048	0.293	0.262	0.038

Note. Robust standard errors in brackets clustered at school level.
[a]Control variables include all the variables in Table 2 and the township dummies.
**Significant at 5%; ***Significant at 1%.

coefficient on the interaction term between the treatment and being a boy is positive (row 2, column 3). As a result, the difference in the treatment effect between the girls and boys during the entire treatment of three semesters is 0.09 standard deviation and insignificant (row 2, column 1). In contrast, boys and girls were affected by the treatments similarly in the fifth grade, where none of the differences across the three time horizons are significant and all the scales are small (row 2, columns 4–6).

Discussion and conclusion

In this paper, we present the results from a randomized field experiment of a CAL program in 72 rural public schools in Ankang Prefecture, Shaanxi Province. The study involves 2741 third grade and fifth grade boarding students. To evaluate the effectiveness of the program, third and fifth grade students in 36 randomly selected schools from the entire sample served received the CAL intervention. Phase I of the program was held

Table 7. Ordinary Least Squares Estimators of the Heterogeneous Effects of the CAL Intervention on Standardized Math Test Scores by Gender

Dependent variable: standardized post-CAL math test score – standardized baseline math test score

		Third grade			Fifth grade		
		Phase I and Phase II (1)	Phase I (2)	Phase II (3)	Phase I and Phase II (4)	Phase I (5)	Phase II (6)
[1]	Treatment (1 = treatment group; 0 = control group)	0.30***	0.29***	0.01	0.29***	0.10	0.19**
		(0.09)	(0.09)	(0.11)	(0.09)	(0.08)	(0.08)
[2]	Treatment × Gender (1 = boy; 0 = girl)	–0.09	–0.21**	0.11	–0.05	0.02	–0.07
		(0.10)	(0.09)	(0.10)	(0.08)	(0.09)	(0.10)
[3]	Control variables[a]	Yes	Yes	Yes	Yes	Yes	Yes
[4]	Observations	1038	1038	1038	1388	1388	1388
[5]	R^2	0.322	0.304	0.049	0.293	0.261	0.038

Note. Robust standard errors in brackets clustered at school level.
[a]Control variables include all the variables in Table 2 and the township dummies.
**Significant at 5%; ***Significant at 1%.

for an intervention period of one semester in 2011. Phase II of the program was implemented for an intervention period of one academic year during 2011–2012. The remaining 36 schools served as control schools. This paper contributes to the limited understanding of whether a CAL program has a persistent long-term impact on student learning.

Our results indicate that the CAL program that was implemented for 1.5 years had significant beneficial effects on students' academic outcomes. Two 40-min CAL math sessions per week for 1.5 years increased students' standardized math scores by 0.25 standard deviation for the third grade boarding students and 0.26 standard deviation for the fifth grade boarding students. In addition, we have shown that the program effect on student gains in math learning is persistent over 1.5 years. In other words, the students continued to improve in math when using the CAL program even after they had become accustomed to using both computers and the software.

Our results suggest that using CAL to complement traditional classes is an effective way to mediate learning activities of students. Although we do not know precisely why, it appears that the remedial tutoring provided by the program to students after regular schooling hours effectively improved students' performance. The animation or game features of the software also appear to stimulate students' interest. By completing the tasks in the game-based exercises without receiving any instruction from the math teacher, students may also have become more engaged and active learners. If this is what is driving the results, it appears as if our study is consistent with the prediction of activity theory and our CAL program is providing more than just an effective piece of software.

The different components of the CAL program, including reviewing daily lessons, answering remedial exercises through game-based activities, and giving feedback to students on the accuracy of their answers, may have all contributed to the positive impact on student learning. The CAL sessions were organized after school when the control students had already gone home. According to other studies, only 10% of students have home computers and there are almost no commercial remedial tutoring resources available to rural students (Lai *et al.*, 2012a; Yang *et al.*, 2013). Therefore, it is not likely that control students were able to engage in game-based remedial tutoring. The control students also were not likely to have been able to review their lessons in a way that was as engaging or interesting as the ways that were provided to the treatment students through the animated CAL software. Therefore, our results appear to be consistent with activity theory. They suggest that by changing the format of learning tasks, adding learning materials, and engaging students in learning and assessing their own progress, the intervention integrated students and their lessons in new ways that boost test scores in the long term (Garris *et al.*, 2002).

Another possible mechanism is that working in pairs has made a difference relative to working individually. Our intervention provides an opportunity to test the hypothesis that working in pairs is better than working individually. When we randomly assigned students in pairs to work during CAL sessions, not all students were able to pair up. This was because some schools had an odd number of students. As a result, approximately 6% of the CAL school students were not assigned a peer. Because of this, our sample included students both with and without a partner. By comparing the improvements in math test scores between the students working in pairs and the students working individually, we cannot reject the null hypothesis that there is no difference in gains in learning. In other words, students in pairs and those working alone appear to have benefitted equally from the CAL treatment. Interested readers can refer to the online file for more details regarding the results (http://reap.fsi.stanford.edu/publications).

The CAL intervention worked similarly well for the poorer performing and better performing students in the third and fifth grades during the entire treatment of three semesters, during Phase I only and during Phase II only. We also found that the third grade girls improved more than the boys in math in the short term.

Interesting questions remain about whether the CAL program can be made more efficient in improving student learning in developing countries. For example, future research can be conducted to explore whether the interaction between the two students who share one computer during the CAL sessions is beneficial or harmful to student learning. Switching control of the computer may reduce the learning time of a single student (Rogers & Lindley, 2004). However, the interaction and discussion between students may improve the efficiency of learning (Stahl *et al.*, 2006). Future studies should be conducted to explore whether interaction or what kind of interaction between students during the CAL classes can help the students gain more in academic performance in rural China. Moreover, studies should also be conducted to explore the impacts of CAL on other key subjects in the national primary school curriculum, such as Chinese and English. Chinese language skills are particularly important because they have been found to affect off-farm work opportunities and wages (Li, Sato, & Sicular, 2013). English test scores have also been found to be one of the indicators that best predict students' chance to college admission and level of post-college income (Li, Meng, Shi, & Wu, 2012). If CAL can be made to effectively improve Chinese and English language skills in addition to math skills among rural students, there will be important policy implications. This will be especially significant if the central government in China fulfils its stated goal of placing computer rooms in every rural school.

References

Angrist, J. D., & Pischke, J. S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton: 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. doi:10.1162/qjec.122.3.1235

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. doi:10.1257/pol.1.1.52

Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, *72*(1), 101–130.

Chang, C.-Y. (2002). Does computer-assisted instruction + problem solving = improved science outcomes? A pioneer study. *The Journal of Educational Research*, *95*(3), 143–150.

China National Bureau of Statistics (CNBS) (2011). *China national statistical yearbook, 2011*. Beijing, China: China State Statistical Press.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730.

Cotton, K. (2002). Computer-assisted instruction. School improvement research series (SIRS). Retrieved April, 10, 2002.

Cuban, L., & Kirkpatrick, H. (1998). Computers make kids smarter–right? *Technos*, *7*(2), 26–31.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Ecalle, J., Magnan, A., & Calmus, C. (2009). Lasting effects on literacy skills with a computer-assisted learning using syllabic units in low-progress readers. *Computers & Education*, *52*(3), 554–561.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*(4), 441–467.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, *2*, 945–1017.

Glewwe, P., Park, A., & Zhao, M. (2011). A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China. J-PAL Working Paper. Retrieved from http://www.e-district.org/userfiles/553/file/better-vision-development.pdf

Gokhale, A. A. (1996). Effectiveness of computer simulation for enhancing higher order thinking. Retrieved from http://www.scholar.lib.vt.edu/ejournals/JITE/v33n4/jite-v33n4.gokhale

Günther, V. K., Schäfer, P., Holzner, B. J., & Kemmler, G. W. (2003). Long-term improvements in cognitive performance through computer-assisted cognitive training: A pilot study in a residential home for older people. *Aging & Mental Health*, *7*(3), 200–206.

He, F., Linden, L., & MacLeod, M. (2008). *How to teach English in India: Testing the relative productivity of instruction methods within the Pratham English language education program*. New York, NY: Columbia University. Mimeographed Document.

Imbens, G. M., & Wooldridge, J. M. (2008). *Recent developments in the econometrics of program evaluation*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14251

Inal, Y., & Cagiltay, K. (2007). Flow experiences of children in an interactive social game environment. *British Journal of Educational Technology*, *38*(3), 455–464.

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2011). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing*. REAP Working Paper.

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2012a). *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. Retrieved from http://www.fsi.fsi.stanford.edu/sites/default/files/CAL_qinghai_aug16_mwb_fang_edcc_clean_sdr.pdf

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2012b). *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. Retrieved from http://www.fsi.fsi.stanford.edu/sites/default/files/CAL_qinghai_aug16_mwb_fang_edcc_clean_sdr.pdf

Li, H., Meng, L., Shi, X., & Wu, B. (2012). Does attending elite colleges pay in China? *Journal of Comparative Economics*, *40*(1), 78–88.

Li, S., Sato, H., & Sicular, T. (2013). *Rising inequality in China: Challenges to a harmonious society*. Cambridge: Cambridge University Press. Retrieved from http://www.books.google.com/books?hl=zh-CN&lr=&id=0dCaAAAAQBAJ&oi=fnd&pg=PR8&dq=Li,+Sato+&+Sicular,+2011&ots = dOsACQ4ngp&sig = mDQZat2WcBwDi1tMUyCY7xj7Eug

Lim, C. P. (2002). A theoretical framework for the study of ICT in schools: A proposal. *British Journal of Educational Technology*, *33*(4), 411–421.

Linden, L. L. (2008). *Complement or substitute?: The effect of technology on student achievement in India*. Working Paper, Columbia University: InfoDev. Retrieved from http://www.povertyactionlab.org/sites/default/files/publications/Complement%20of%20Substitute-%20The%20Effect%20of%20Technology%20on%20Student%20Achievement%20in%20India.pdf

Luo, R., Shi, Y., Zhang, L., Liu, C., Rozelle, S., Sharbono, B., . . . Martorell, R. (2012). *Nutrition and educational performance in rural China's elementary schools: Results of a randomized control trial in Shaanxi Province*. Rochester, NY: Social Science Research Network. Retrieved from http://www.papers.ssrn.com/abstract=2246923. SSRN Scholarly Paper No. ID 2246923.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude. Learning, and Instruction*, *3*(1987), 223–253.

Marjanovic, O. (1999). Learning and teaching in a synchronous collaborative environment. *Journal of Computer Assisted Learning*, *15*(2), 129–138.

McEwan, P. J. (2013). *Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments*.

Ministry of Education. 2012. *12th Five-Year Plan for Integrating Information Technology into Education*. Retrieved from http://www.edu.cn/zong_he_870/20120723/t20120723_813704_50.shtml

Mo, D., Yi, H., Zhang, L., Shi, Y., Rozelle, S., & Medina, A. (2012). Transfer paths and academic performance: The primary school merger program in China. *International Journal of Educational Development*, *32*(3), 423–431.

Mosteller, F. (1997). The Tennessee study of class size in the early school grades. *Bulletin of the American Academy of Arts and Sciences*, *50*(7), 14–25.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, *24*(2), 129–144.

Ong, C.-S., & Lai, J.-Y. (2006). Gender differences in perceptions and relationships among dominants of e-learning acceptance. *Computers in Human Behavior*, *22*(5), 816–829.

Rico García, M., & Vinagre Arias, F. (2000). A comparative study in motivation and learning through print-oriented and computer-oriented tests. *Computer Assisted Language Learning*, *13*(4–5), 457–465.

Roesch, A., Gruber, H., Hawelka, B., Hamm, H., Arnold, N., Popal, H., . . . Stolz, W. (2003). Computer assisted learning in medicine: A long-term evaluation of the 'Practical Training Programme Dermatology 2000'. *Informatics for Health and Social Care*, *28*(3), 147–159.

Rogers, Y., & Lindley, S. (2004). Collaborating around vertical and horizontal large interactive displays: Which way is best? *Interacting With Computers*, *16*(6), 1133–1152.

Salamonson, Y., & Andrew, S. (2006). Academic performance in nursing students: Influence of part-time employment, age and ethnicity. *Journal of Advanced Nursing*, *55*(3), 342–349.

Schaefer, S., & Warren, J. (2004). Teaching computer game design and construction. *Computer-Aided Design*, *36*(14), 1501–1510.

Squires, D., & McDougall, A. (1994). *Choosing and using educational software: A teachers' guide*. Bristol: Psychology Press. Retrieved from http://www.books.google.com/books?hl=zh-CN&lr=&id=HxSHVkB9epMC&oi=fnd&pg=PR8&dq=mcdougall+1994+evaluation+framework+computer+software&ots=vYfJoYZu3l&sig=mPauL-ksANZ-jAqYqne923Q4DaA

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: A historical perspective. Cambridge Handbook of the Learning Sciences, 2006. Retrieved from http://www.gerrystahl.net/cscl/CSCL_English.htm

Szabo, M., & Poohkay, B. (1996). An experimental study of animation, mathematics achievement, and attitude toward computer-assisted instruction. *Journal of Research on Computing in Education*, *28*(3), 390–402.

Tondeur, J., Van Braak, J., & Valcke, M. (2007). Towards a typology of computer use in primary education. *Journal of Computer Assisted Learning*, *23*(3), 197–206.

Tüzün, H., Yılmaz-Soylu, M., Karakuş, T., İnal, Y., & Kızılkaya, G. (2009). The effects of computer games on primary school students' achievement and motivation in geography learning. *Computers & Education*, *52*(1), 68–77.

Woessmann, L. (2004). How equal are educational opportunities? Family background and student achievement in Europe and the US. Retrieved from http://www.papers.ssrn.com/sol3/papers.cfm?abstract_id=528209

Yang, Y., Hu, X., Qu, Q., Lai, F., Shi, Y., Boswell, M., & Rozelle, S. (2013). Roots of tomorrow's digital divide: Documenting computer use and internet access in China's elementary schools today. *China & World Economy*, *21*(3), 61–79.

Zhao, M., & Glewwe, P. (2010). What determines basic school attainment in developing countries? Evidence from rural China. *Economics of Education Review*, *29*(3), 451–460.

Appendix I

Students' Interest Rating in the CAL Software by the End of Phase II.

			Interest rating by computer experience				Interest rating by baseline math score			
		Mean interest rating of the total sample	Students who had used computer before the program (2) Mean	Students who had never used computer before the program (3) Mean	Difference = (2)-(3)	*p* value	Students with baseline math score higher than the median (6) Mean	Students with baseline math score lower than or equal to the median (7) Mean	Difference = (6)-(7)	*p* value
		(1)	(SD)	(SD)	(4)	(5)	(SD)	(SD)	(8)	(9)
[1]	Third grade	88.28	88.09 (19.34)	89.33 (18.64)	−1.24	0.60	87.29 (21.63)	88.91 (17.53)	−1.63	0.35
[2]	Fifth grade	83.39	83.48 (18.28)	82.34 (18.82)	1.14	0.68	82.39 (20.43)	84.27 (16.20)	−1.88	0.20