

# **Institutions, Implementation, and Program Effectiveness: Evidence from a Randomized Evaluation of Computer-Assisted Learning in Rural China**

**Di Mo, Yu Bai, Yaojiang Shi, Cody Abbey, Linxiu Zhang, Scott Rozelle, Prashant Loyalka**

## **Abstract**

There is limited evidence on the degree to which differences in implementation among institutions matter for program effectiveness. To examine this question, we conducted an experiment in rural China in which public schools were randomly assigned to one of three treatments: a computer-assisted learning program (CAL) implemented by a government agency, the same program implemented by an NGO, and a pure control. Results show that compared to the pure control condition and unlike the NGO program, the government program did not improve student achievement. Analyzing impacts along the causal chain, we find that government officials were more likely to substitute CAL for regular instruction (contrary to protocol) and less likely to directly monitor program progress. Correlational analyses suggest that these differences in program implementation were responsible for the lack of impacts.

**Keywords:** Computer-assisted learning; education; external validity of program; implementation by government; randomized controlled trial

---

**Working Paper**

**April 2020**

**[reap.fsi.stanford.edu](http://reap.fsi.stanford.edu)**



**Institutions, Implementation, and Program Effectiveness: Evidence from a Randomized  
Evaluation of Computer-Assisted Learning in Rural China**

Di Mo<sup>a</sup>, Yu Bai<sup>b,\*</sup>, Yaojiang Shi<sup>c</sup>, Cody Abbey<sup>a</sup>, Linxiu Zhang<sup>d, e</sup>, Scott Rozelle<sup>a</sup>, and Prashant Loyalka<sup>a</sup>

<sup>a</sup> Rural Education Action Program (REAP), Food Security and the Environment (FSE), Freeman Spogli Institute (FSI), Stanford University, Stanford, USA

<sup>b</sup> School of Economics, Minzu University of China, Beijing, China

<sup>c</sup> Center for Experimental Economics in Education (CEEE), Shaanxi Normal University, Xi'an, China

<sup>d</sup> Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

<sup>e</sup> UNEP-International Ecosystem Management Partnership, Beijing 100101, China, Beijing, China

\* Corresponding author:

Yu Bai

School of Economics, Minzu University of China  
27 Zhongguancun South Avenue Beijing, China, 100081

Email: yubai@muc.edu.cn

Phone: 86-186-0040-5923

Email of authors:

Di Mo (dimo@stanford.edu)

Yu Bai (yubai@muc.edu.cn)

Yaojiang Shi (shiyaojiang7@gmail.com)

Cody Abbey (cjabbey@stanford.edu)

Linxiu Zhang (lxzhang.ccap@igsnr.ac.cn)

Scott Rozelle (rozelle@stanford.edu)

Prashant Loyalka (loyalka@stanford.edu)

# **Institutions, Implementation, and Program Effectiveness: Evidence from a Randomized Evaluation of Computer-Assisted Learning in Rural China**

## **Abstract**

There is limited evidence on the degree to which differences in implementation among institutions matter for program effectiveness. To examine this question, we conducted an experiment in rural China in which public schools were randomly assigned to one of three treatments: a computer-assisted learning program (CAL) implemented by a government agency, the same program implemented by an NGO, and a pure control. Results show that compared to the pure control condition and unlike the NGO program, the government program did not improve student achievement. Analyzing impacts along the causal chain, we find that government officials were more likely to substitute CAL for regular instruction (contrary to protocol) and less likely to directly monitor program progress. Correlational analyses suggest that these differences in program implementation were responsible for the lack of impacts.

**Keywords:** Computer-assisted learning; education; external validity of program; implementation by government; randomized controlled trial

**JEL Classification:** I24, I28, O33

# **Institutions, Implementation, and Program Effectiveness: Evidence from a Randomized Evaluation of Computer-Assisted Learning in Rural China**

## **1 Introduction**

Integrating technology into education has been found to be an effective means of helping students across the world (Pal 2006; Banerjee et al. 2007; Ebner and Holzinger 2007; Lai and Gu 2011; Escueta et al. 2017). Studies from developing countries, such as India, find that computer-assisted learning (CAL) programs can compensate for teacher shortages or poor teaching quality and ultimately improve student math scores (Pal 2006; Banerjee et al. 2007; Muralidharan et al. 2017a). In rural China, Mo et al. (2015) and Bai et al. (2016) demonstrate that CAL programs have positive and statistically significant impacts on student achievement in math, language arts, and English in both the short and long term. Escueta et al. (2017), which provides a literature review of studies on the effectiveness of CAL programs, finds that the majority of studies (21 out of 29) find positive and statistically significant effects on student learning in both developed and developing countries.

Although introducing technology into the formal education system may be a cost-effective way to improve student outcomes in developing countries, technology programs may fail to produce impacts because of poor implementation. In particular, the nature of principal-agent relationships may undermine program effectiveness. For example, education officials (principals) may not devote sufficient resources to help school administrators and teachers (agents) implement or upscale a program (World Bank 2003). The actions of agents may also be difficult to monitor. Schools are often far away from the central offices of educational

authorities. As a result, it is often difficult to know whether school administrators and teachers have set aside time to carry out the program with sufficient frequency, regularity, and quality.

This paper contributes to the literature by comparing program implementation and effectiveness under different types of institutions. To this end, we have three specific objectives. First, we measure the impact on student outcomes of a CAL program that is implemented by a government agency (henceforth, the government CAL program) versus a pure control condition. Second, we examine the impact of an identical CAL program implemented by an NGO (henceforth the NGO CAL program) versus the pure control and the government CAL program. Third, we explore mechanisms—including differences in implementation across the treatment groups—that may underlie potential differences in student outcomes.

To meet these objectives, we conducted a cluster-RCT in rural China. We selected 120 primary schools in one of China's poorest provinces and then randomized the schools into three groups: schools that received a CAL program for English language learning that was implemented by a government agency (henceforth, government schools), schools that received an identical CAL program that was implemented by an NGO (henceforth, NGO schools), and schools that did not receive any CAL program (henceforth, control schools).<sup>1</sup> We assessed the program's effectiveness by measuring how the 5,574 Grade 4 students in the study scored on a standardized English language test after one academic year of the CAL program, holding constant the scores on a similar standardized test given during the baseline. To understand potential mechanisms that led to any differences in effects, we measured the effort that

---

<sup>1</sup> Researchers are concerned that governments (compared to non-governmental entities) may not devote equally sufficient resources towards program implementation (Deaton 2010; Banerjee et al. 2017; Muralidharan and Niehaus 2017). However, we know of few experimental evaluations that directly compare the same program implemented by different types of institutions. Indirect comparisons abound (see the discussion in Banerjee et al. 2017). For example, a non-experimental meta-analysis by Vivaldi (forthcoming) shows larger effect sizes when programs are implemented by NGOs as opposed to governments.

principals—education bureau personnel (i.e., government officials) and NGO staff—put into program implementation and monitoring. We also assessed the extent to which agents (school administrators and teachers) adhered to implementation protocols. Finally, we conducted correlational analyses to explore the degree to which different mechanisms were associated with differences in students’ academic outcomes.

Our findings, when taken together, indicate that differences in implementation among institutions matter for program effectiveness. In regard to direct impacts, we find that the government CAL program was ineffective, relative to the control group, at improving student outcomes. By contrast, the NGO CAL program resulted in improved student outcomes relative to either the control group or the government CAL program. In addition, we find that certain measures of the fidelity of the implementation process (what we call “programmatic content” for short), did not differ between the two program intervention conditions. For example, the quality of teacher training associated with the CAL program did not differ between the government and NGO treatment arms. In addition, the number and frequency of CAL sessions held in the government and NGO schools were the same.

Although programmatic content did not differ, we do find other potentially critical differences in the degree to which agents (schools and teachers) adhered to the CAL program protocol. For example, according to protocol, the CAL program was supposed to be held during computer class sessions instead of during English class and run by computer teachers instead of English teachers. Ultimately, however, government schools were more likely to violate these aspects of the protocol. Compared to NGO schools, government schools were 29 percentage points more likely (43 percent of government schools versus 14 percent of NGO schools) to substitute English classes with the CAL program. This resulted in students in the government

CAL program receiving less teacher-led instruction overall and, therefore, less English instruction overall than students in the NGO CAL program. Government schools were also 23 percentage points more likely (35 percent of government schools versus 12 percent of NGO schools) to assign English teachers to run the CAL program.

We also find substantial differences in monitoring across the two intervention groups. In fact, our results indicate that government officials were less likely to directly monitor the schools (call or visit the schools to follow up with program progress) than were the NGO staff. Government officials also were much less likely to help program teachers solve technical problems.

Finally, simple descriptive results from correlational analyses suggest that both the substitution of regular instruction and lack of direct monitoring may have led to the lack of statistically significant impacts in government schools. While we do not have data to determine the reasons behind the increased substitution and lack of monitoring, we posit that the government may have faced greater resource constraints or less pressure from outside stakeholders than the NGO. Regardless of the reasons, the results suggest that not addressing resource constraints to avoid substitution and a lack of direct monitoring were two key institutional features that hampered program effectiveness.

Taken together, the results of our study contribute to a small but growing literature on the role of institutions in program roll out and scale up (Banerjee et al. 2017). Bold et al. (2013) also looks at an education program implemented via both government and NGO.<sup>2</sup> The study finds that when a contract teacher program was implemented by an NGO, the program yielded positive

---

<sup>2</sup> Although Bold et al. (2013) conducts a direct experimental comparison of NGO- versus government-led programs, this comparison was also overlaid with three additional treatment comparisons. Differences in effects between the NGO- and government-led treatment arms could therefore have been due to potential complementarities between the various treatment regimes (Mbiti et al. 2019).

and statistically significant effects on student achievement. In contrast, the treatment effects were indistinguishable from zero in schools that received contract teachers from the local government. The authors conclude that inefficient implementation and monitoring by local officials may have led to lower levels of teacher effort and, ultimately, a lack of program effectiveness. Although not directly comparing program effectiveness and implementation between different institutions, Banerjee et al. (2017) also shows that a teaching training intervention in India could have been effective if officials had provided sufficient mentoring and monitoring. Taken together, the results of these two studies, as well as the present study, suggest that more complex educational programs (such as teacher recruitment, teacher training, and introducing a computer-assisted learning program into schools) may require additional institutional support that may not necessarily be required for less complex programs.

## **2 Sampling, Data, and Methods**

### **2.1 Sampling and the Process of Randomization**

We conducted a clustered RCT of CAL in rural schools in northwest China during the 2013–2014 academic year. A total of 120 primary schools, comprising 5,574 Grade 4 students in poor minority areas in China’s Qinghai Province, were included in our study. We focus on Grade 4 students because the CAL program provided remedial tutoring for the subject of English, which many students in Qinghai begin to study as early as the fourth grade (Zheng 2009). The pace of the CAL program was matched to the English teaching curriculum in the sample schools.

We followed three steps to select the sample (Figure 1). First, to focus our study on students from poor rural areas, we restricted our sampling frame to Haidong Prefecture, a poor minority area in Qinghai Province in northwest China. Among the 31 provinces in mainland



China, Qinghai ranked 30th in terms of GDP per capita in 2013 (National Bureau of Statistics 2014). The annual per capita net income of the selected prefecture was only 6,150 RMB (approximately 990 USD; National Bureau of Statistics 2014). Second, after choosing Haidong Prefecture in Qinghai Province, we selected the sample counties. All six counties in the prefecture were chosen to be included in our sampling frame. Five of the counties are nationally designated poor counties (National Bureau of Statistics 2014).

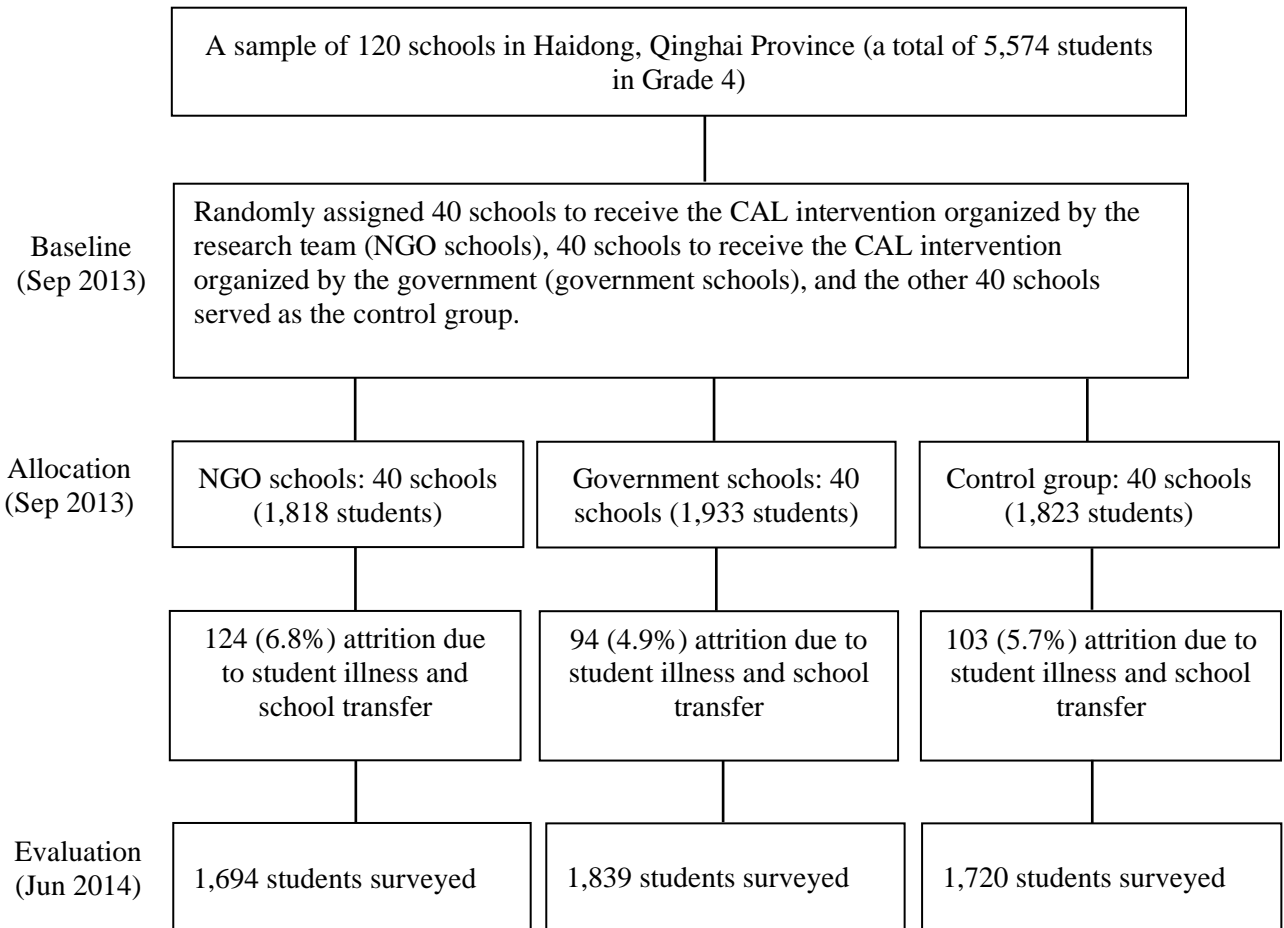


Figure 1: Experimental Profile

Third, we selected the sample schools based on our power calculation and the availability of school computer facilities. According to our power calculation, we needed 40 schools per treatment arm to find a minimum detectable effect of 0.20 standard deviations (SDs) in the outcome test scores, with a power of 0.80, a 5 percent significance level, an intra-cluster correlation (ICC) of 0.1, a pre- and post-intervention correlation of 0.40, and an average of 50 observations in each school (cluster). To generate a sampling frame to choose 120 schools, we obtained a comprehensive list of schools in the six counties from each county's local education bureau. We then restricted our sample to schools that met the minimum requirement of the CAL

program for computer facilities. The minimum requirement was a ratio of one computer to four students (or 0.25). This ratio was needed to ensure that, at most, two students would be able to share one computer. We found a total of 130 schools in the six sample counties that met the requirement, and then we included 120 schools in the experiment.

Next, we randomly assigned the participating schools to one of three groups: the NGO treatment arm, the government treatment arm, or the control (Figure 1). Ultimately, we randomly assigned 40 schools to receive a CAL treatment that was implemented by the NGO (*NGO schools*) and another 40 schools to receive a CAL treatment that was implemented by the government officials (the staff of the prefectural/county bureaus of education – the lowest levels of the education hierarchy – and the school administrators and teachers in the county’s treatment schools; *government schools*). The remaining 40 schools, all of which had computer rooms, were assigned to the control group (*control schools*) and, as such, did not participate in the CAL program. Ultimately, there were a total of 1,818 students in the 40 NGO schools, 1,933 students in the 40 government schools, and 1,823 students in the 40 control schools.

Our randomization process created a balanced sample across the treatment and control groups. To analyze whether there were any statistically significant differences among the three groups, we used a set of student characteristics. In doing so, we regressed the baseline characteristics on the treatment dummy variables. The results show that none of the variables exhibit statistically significant differences among the three groups (Table 1). In addition, all of the differences between treatment and control groups are small in magnitude.

Table 1. Comparison of the baseline characteristics of the treatment and control groups prior to attrition

		NGO schools		Government schools		P-value for joint F test
		Coefficient	SE	Coefficient	SE	
		(1)	(2)	(3)	(4)	
[1]	Student baseline English score ( <i>SD</i> )	-0.02	(0.22)	0.03	(0.18)	0.96
[2]	Student gender (1 = male, 0 = female)	-0.01	(0.02)	-0.00	(0.02)	0.71
[3]	Student age (year)	0.04	(0.10)	-0.07	(0.10)	0.60
[4]	Student ethnic minority (1 = yes, 0 = no)	-0.00	(0.09)	-0.01	(0.08)	0.99
[5]	Student self-efficacy scale (0-4 pts)	0.00	(0.03)	0.01	(0.03)	0.94
[6]	Used computer before (1 = yes, 0 = no)	-0.05	(0.07)	-0.00	(0.07)	0.70
[7]	Only child (1 = yes, 0 = no)	0.02	(0.03)	0.02	(0.03)	0.78
[8]	Mom junior high or higher (1= yes, 0= no)	-0.04	(0.06)	0.03	(0.05)	0.42
[9]	Dad junior high or higher (1 = yes, 0 = no)	0.01	(0.04)	0.03	(0.04)	0.67
[10]	Mom has migrant job (1 = yes, 0 = no)	-0.03	(0.03)	-0.02	(0.03)	0.51
[11]	Dad has migrant job (1 = yes, 0 = no)	-0.03	(0.03)	-0.04	(0.03)	0.31

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

Although the sample included 5,574 Grade 4 students at the time of the baseline survey, there was 5.8 percent attrition from the survey by the end of the study. This attrition was due to a number of reasons. The most common reasons were school transfers and extended absences due to illness or injuries (Figure 1). In the end, we were able to follow up with 5,253 students.

To understand the characteristics of those who dropped out of the study and to assess whether attrition affected the validity of the randomization, we regressed attrition status on the treatment variables. A comparison of the attrition rates among the two treatment groups and the control group showed that the attrition rates were not related to treatment status (Table 2).

Table 2. Comparison of attrition between the treatment and control groups

		Dependent variable: Attrition (1 = yes, 0 = no)
		(1)
[1]	NGO schools (1 = yes, 0 = no)	0.01 (0.02)
[2]	Government schools (1 = yes, 0 = no)	-0.01 (0.01)
[3]	Constant	0.06*** (0.01)
[4]	Observations	5,574
[5]	R-squared	0.00

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

We also used the sample that completed the study (i.e., students who took part in both the baseline and post-intervention surveys and were included in the final analytic sample) and regressed the baseline characteristics on treatment variables. None of the regressions shows statistically significant differences between the treatment and control groups (Table 3). In other words, our results show that student characteristics were well balanced between the treatment and control groups, both prior to and after attrition.

Table 3. Comparison of baseline characteristics of the treatment and control groups after attrition

		NGO schools		Government schools		P-value for joint F test
		Coefficient	SE	Coefficient	SE	
		(1)	(2)	(5)	(4)	
[1]	Student baseline English score ( <i>SD</i> )	0.01	(0.22)	0.03	(0.18)	0.99
[2]	Student gender (1 = male, 0 = female)	-0.01	(0.02)	0.00	(0.02)	0.87
[3]	Student age (year)	0.03	(0.10)	-0.06	(0.10)	0.66
[4]	Student ethnic minority (1 = yes, 0 = no)	-0.01	(0.09)	-0.01	(0.08)	0.99
[5]	Student self-efficacy scale (0-4 pts)	-0.00	(0.03)	0.01	(0.03)	0.90
[6]	Used computer before (1 = yes, 0 = no)	-0.05	(0.07)	-0.00	(0.07)	0.72
[7]	Only child (1 = yes, 0 = no)	0.01	(0.03)	0.02	(0.03)	0.76
[8]	Mom junior high or higher (1=yes, 0=no)	-0.03	(0.06)	0.03	(0.05)	0.45
[9]	Dad junior high or higher (1=yes, 0=no)	0.01	(0.04)	0.03	(0.04)	0.77
[10]	Mom has migrant job (1=yes, 0=no)	-0.02	(0.03)	-0.02	(0.03)	0.66
[11]	Dad has migrant job (1=yes, 0= no)	-0.03	(0.03)	-0.05	(0.03)	0.32

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

## 2.2 Experiment Arms/Interventions

Both the intervention organized by the education bureau and the intervention organized by the NGO involved CAL programs that used computers and sets of remedial tutoring software to complement the regular English curriculum of the students. The education bureau is a formal part of the prefectural government and has the responsibility of managing school-based programs, among other responsibilities. The NGO is a university-based, education-oriented entity that is involved, throughout western China, in implementing projects and programs, mainly in rural schools.

In both the government and NGO schools, the protocol was to implement the CAL program during the period of time at school when students had a formal computer course scheduled.<sup>3</sup> Before the CAL program, fourth graders in each of the sample schools had one computer class that met twice per week. During the computer classes, the students would use computers to learn basic computer operations, such as how to use a mouse or the keyboard. No

<sup>3</sup> According to our investigations prior to the study, the quantity and quality of the usual computer classes were low; as such, CAL sessions did not replace any substantial learning in which the students otherwise might have engaged.

learning in any subjects other than basic computer skills took place during computer classes.<sup>4</sup> The CAL program protocol required the CAL sessions to be organized during each school's computer class time period, that is, during the two 40-minute program sessions each week. To facilitate the sessions' being held during computer classes, the protocol also required computer teachers, and not English teachers, to be assigned to organize the CAL program and to supervise the session.

To facilitate the implementation of the CAL protocol, the NGO and government agency compensated the teachers who instructed the CAL sessions in their respective treatment arms with a stipend of 500 RMB (approximately 80 USD) every semester. The stipend was distributed to the program teachers at the end of the 2013–2014 academic year only if they faithfully implemented the CAL program. To monitor how closely the teachers followed the protocol, the NGO recruited enumerators from universities in Haidong Prefecture to visit the CAL schools during the program period. The enumerators randomly selected students from each treatment class and surveyed them about how the CAL classes were organized. Instead of recruiting enumerators, the government agency sent an official from the county bureau of education, the county program manager, to the schools for monitoring.

In both sets of the treatment schools, two pieces of software (henceforth, the CAL software package) were used. One set of software for the program was designed to improve the basic competencies of students in the uniform national English curriculum. This software,

---

<sup>4</sup> Although the Central Ministry of Education sets the national curriculum for every subject, including computer classes, we found that many rural Qinghai schools did not adhere to this curriculum. Our data indicated that only 52 percent of the students reported that they had ever had a computer class, and 78 percent of these reported they had only one 40-minute computer class (instead of two 40-minute classes) per week. Similar results were found in Xi et al. (2017). As far as curriculum, when we called sample schools that held computer classes regularly, we found that all computer teachers reported that they taught only basic operations, such as how to use the mouse or keyboard. In fact, students also could learn such skills in CAL sessions, which require typing and mouse operation to use the software. The CAL teachers also are required to provide students with guidance in this regard.

obtained from a commercial IT company, provided both animated reviews of the lesson of the day and game-based remedial exercises in English (for that lesson) for the Grade 4 students.

The NGO paid to have other software developed. The second software package provided a large number of additional remedial exercise questions. The NGO worked with teachers and experts from the Beijing-based Teaching English to Speakers of Other Languages (TESOL) organization to choose the questions. The questions were then integrated into an animation-based game interface to make it interesting and fun for the students.

Although the CAL software was based on the material that students learned in class and did not impart any substantial new material, it had several features that may have uniquely contributed to student learning. First, the animated and game-based nature of the types of software may have increased student interest and motivation in learning the subject area. Second, the software provided one-on-one remedial tutoring with timely feedback. Once a student answered a question incorrectly, a window appeared with the correct answer and a detailed explanation. To prevent students from proceeding to the next question without digesting the explanation, the window could not be closed until ten seconds had passed. Thus, although based on the same curriculum, the software included several elements that traditional classroom pedagogy does not possess and that could potentially improve student learning outcomes.

English was chosen as the program subject for several reasons. First, English is one of the main subjects used to test students as part of the competitive exam system in China, which allows students to compete for positions in high school and college (Bolton and Graddol 2012; McKay 2002). In fact, at the time of the survey, English represented approximately one-third of the total points in both the high school and college entrance exams. Second, English teaching and English learning are particularly weak in the poor rural areas of China (Hu 2005; Li 2002; Zhao



2003). Studies have shown that a low English score is one of the primary factors that keeps rural students from attending high school in China (Loyalka et al. 2014). According to the literature, English teachers are of notoriously poor quality in rural China (Hu 2005; Hu 2009). Based on these reasons, English learning was the targeted subject of the CAL program.

Our data show that the quality of the English teaching was particularly poor in rural Qinghai. Although primary school students started to learn English in Grade 3, and 80 percent of fifth graders had four to five 40-minute English classes each week, according to the reports from students and school administrators of sample schools, they lacked qualified teachers, and local public schools lacked English teachers in general. The student-English teacher ratio was 178:1, whereas the average student-teacher ratio in rural Qinghai primary school was 21:1 (Ministry of Education, 2014). Nearly 53 percent of our sample schools had no more than two English teachers per school, and more than one-third of the sample schools had only one English teacher. Furthermore, the quality of English teachers was poor. Only 34 percent of the English teachers had a college education, and 33 percent had only a vocational high school education.

### **2.2.1 NGO schools**

The NGO implemented the CAL program in the 40 NGO schools. To organize the program for the teachers in these schools, the NGO carefully designed and compiled a CAL implementation protocol. The protocol included a lesson-by-lesson curriculum, a software introduction, and a set of instructions in regard to teacher responsibilities. During a one-day intensive training before the program was launched, the NGO trained the computer teachers on the program protocol. During the program implementation, teachers were called periodically to ensure that the program progressed as planned and the protocol was being followed. The NGO

also conducted random checks of all schools by sending volunteers to each school. Whenever needed, the NGO provided technical support.

The protocol also required computer teachers to organize CAL sessions twice a week during computer classes. In a typical CAL session, the students were supervised to use the software package to learn English on their own. Students watched the animated video that reviewed the materials on which they were receiving instruction during their regular English class sessions (during the previous week or during a time earlier that day). Then, students played games to practice the skills introduced in the video lecture. No instruction on English was supposed to be given during the classes.

### **2.2.2 Government schools**

Among the schools in the government school group, the CAL program was implemented by the prefecture education bureau in a centralized way, the way that most other government programs are carried out. The prefecture education bureau first sent out an official document (directive) to the county education bureau to inform them to launch the CAL program. The directive required officials from the county education bureau to assign an individual, i.e., a county program manager, to be in charge of carrying out the program. The county program managers are officers of the local education bureaus and have daily responsibilities that include mainly carrying out countywide pedagogical research and teaching quality evaluations. In addition to these regular duties, they must carry out tasks stipulated in official directives issued by the provincial government, such as the official directive mentioned above. Based on the directive, it became a core responsibility of the program managers to implement and monitor the CAL program.

The prefecture education bureau then organized a staff training in which all the county program managers were taught about the requirements of the program implementation. After the training, the program managers went back to their own county and organized teacher training programs, conducted with the teachers from government schools at their respective county seats. During program implementation, the county program manager also was responsible for monitoring the program's progress and reporting to the prefecture education bureau.

To help the prefecture bureau to implement the overall CAL program, a team from the NGO was invited as outside experts to train the program managers in the program's protocol. This reliance on outside experts is a common occurrence when new programs are promoted in China's school system. To ensure that the basic program in the government schools was the same as that being implemented in the NGO schools, the NGO expert group prepared identical training materials for both groups, intended to minimize the differences that might exist in the training of teachers in the two types of schools.

The training materials included training slides that covered all of the knowledge that teachers needed for successfully implementing the program. Also included were documents/PowerPoint presentations that summarized the most important features of the software and the program's protocol. All of the training materials were supported by well-tested training videos that the county program managers were required to use to demonstrate the CAL program's implementation protocol to the teachers.

### **2.2.3 Control schools**

As noted, a total of 1,720 Grade 4 students in 40 control schools constituted the control group. During the program, students in the control group did not receive any intervention. To avoid any spillover effects, administrators in the control schools, which were located in school

districts separate from those of the treatment schools, were blinded to the experimental nature of the study. To our knowledge, none of the school administrators, teachers, or students in the control school was aware of the program.

### **2.3 Data Collection**

We conducted two rounds of surveys in the 120 sample schools. The first round was a baseline survey that took place in September 2013 at the beginning of the fall semester. The baseline survey was completed before any implementation of our experiment had begun. The second round of the survey was an evaluation, which was conducted at the conclusion of the study in June 2014, a time that coincided with the end of the spring semester of the 2013–2014 academic year.

In each survey round, the enumeration team visited each sample school and conducted a three-part survey. The first part comprised a 30-minute standardized English test. All of the questions in the English test in the post-intervention survey were different from the questions in the baseline survey. We included only questions that did not overlap with the exercises in the CAL software package. Our enumeration team strictly enforced time limits and proctored the examinations. We normalized the baseline and post-intervention test scores relative to the distribution of the baseline and post-intervention test scores of the control group to ensure that test scores from different rounds of surveying were comparable. We used the normalized English scores as the main measure of the program outcome of English academic performance.

In the second part of the survey, enumerators collected information on the characteristics of students and their families. Based on this part of the survey, we were able to construct a set of demographic and socioeconomic indicators. The dataset included measures of *student gender*, *student age*, whether the *student belongs to an ethnic minority*, whether the student *used a*

*computer before*, whether the student is the *only child* in the family, whether the *mother finished junior high or higher education*, whether the *father finished junior high or higher education*, whether the *mother has a migrant job*, and whether the *father has a migrant job*.

In the second part of the survey, we also collected information on student non-cognitive traits. We adopted the General Self-Efficacy Scale (GSE) to measure *self-efficacy* (Jerusalem and Schwarzer 1992). We collected information on student perceptions of their English teachers and English classes. The measures include *liking of the English teacher*, *liking of the English class*, and the *feedback students received in English classes*. The measure of the feedback that students received in the English classes was generated using five questions about students' interaction with the English teacher in the English classes. These questions reflect the amount of attention that each student perceived to have received from the English teacher.

In the third part of the survey, we gathered information on program implementation. First, we evaluated the quality of the teacher training by administering a 20-minute test to all of the teachers who participated in the teacher training organized by the NGO (in the NGO schools) and the county program managers (in the government schools). The test was designed to evaluate teacher knowledge of the software and program protocol.

In the third section of the evaluation survey, we also documented how well the program protocol was followed by the school administrators and teachers. For example, we documented the number of months that the CAL program was run and how often the CAL sessions were organized during the academic year. Both students and teachers were asked these questions for verification of the information. We also collected information on the time slots during which the CAL sessions took place and which teacher was assigned to organize/supervise the CAL sessions. Finally, we asked teachers whether anyone from the government called or visited the

schools to follow up with, supervise, or monitor program implementation. We also kept track of the interactions between the NGO and the schools by recording whether and how often the NGO called or visited the schools.

## 2.4 Statistical Methods

To estimate the average treatment effect of the CAL program that was implemented by the government and the NGO, we adopted both an unadjusted and adjusted ordinary least squares (OLS) regression analytical approach. The unadjusted OLS analysis regressed the post-intervention outcome variables (e.g., *standardized post-intervention English test score*) on the dummy variables of the treatment status. The model (henceforth, the unadjusted model) we estimated is:

$$y_{is} = \alpha + \beta_1 treat_{1s} + \beta_2 treat_{2s} + \theta y_{0is} + \varepsilon_{is} \quad (1)$$

where  $y_{is}$  is the post-intervention outcome variable for child  $i$  in school  $s$ ,  $treat_{1s}$  is the dummy variable that indicates a student in an NGO school (equal to 1 for students in the NGO schools and 0 otherwise), and  $treat_{2s}$  is the dummy variable that indicates a student in a government school (equal to 1 for students in the government schools and 0 otherwise). Because the treatment was randomly assigned,  $\beta_1$  and  $\beta_2$  in equation (1) provide unbiased estimates of the average treatment effect of the CAL program that was conducted by the NGO personnel and the officials, including the county program managers, from the government. We also included  $y_{0is}$  on the right-hand side, which is the outcome variable at the baseline.

To improve the efficiency of estimation, we also estimated an *adjusted model* by including a set of control variables:

$$y_{is} = \alpha + \beta_1 treat_{1s} + \beta_2 treat_{2s} + \theta y_{0is} + X_{is}\gamma + \varepsilon_{is} \quad (2)$$

where  $X_{is}$  includes a vector of control variables of student and family characteristics (Rows 2 to 11 in Table 1). In all regressions, we accounted for the clustered nature of our sample by constructing Huber-White standard errors, corrected for school-level clustering.

### 3 Results

The data indicate that there is a statistically significant difference between the impact of the government CAL program and the NGO CAL program (Table 4). The NGO CAL program improved student English scores by 0.16 SD (statistically significant at the 5 percent level) if we use equation (1), the unadjusted model (Column 1, Row 1). When using the adjusted model in equation (2), which added control variables, the NGO CAL program improved English scores by 0.18 SD (statistically significant at the 1 percent level; Column 2, Row 1). The signs and the magnitudes (0.16 to 0.18 SDs) of the NGO CAL program are similar to those (from 0.13 to 0.16 SDs) that were generated by other CAL programs in China, which were all implemented by NGOs (Mo et al. 2015).

Table 4. Ordinary least squares estimates of treatment effect on standardized English test scores

		Dependent variable: Evaluation English score (SD)	
		(1)	(2)
[1]	NGO schools (1 = yes, 0 = no)	0.16** (0.07)	0.16** (0.07)
[2]	Government schools (1 = yes, 0 = no)	-0.07 (0.07)	-0.07 (0.07)
[3]	Control variables <sup>a</sup>	No	Yes
[4]	Observations	5,253	5,253
[5]	R-squared	0.64	0.65

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

---

<sup>a</sup> Control variables include the variables shown in Table 1.

In contrast, the program that was implemented by the government was ineffective. The scores of the students in the government CAL program did not improve relative to those of the students in the control schools (Table 4, Row 2). The estimate of the program effect is negative, but not statistically significant (Column 1, Row 2), if we use the unadjusted model in equation (1). When using the adjusted model (equation 2), the effect of the program on standardized English scores is still less than zero (-0.07) and is not statistically significant (Column 2, Row 2). To better understand the magnitude of the impact, we also run a regression based on equation 2, using raw scores as the dependent variable. The results show that the NGO CAL program improved English raw scores by 2.13 points (on a 100-point exam), while the coefficient on government CAL program treatment variable was close to zero (-0.90 points) and not statistically significant (Appendix A).

#### **4 Mechanism analysis**

In this section, we empirically examine why the two identical CAL protocols (government and NGO) produced such different outcomes. We focus on whether the two CAL programs were actually conducted in different ways. In addition to comparing program implementation and monitoring between the two groups, we also examine correlations between these potential causal mechanisms and student achievement. Although potentially illuminating, we approach the correlational analyses with caution, as the mechanisms are clearly not exogenous.

According to our analysis, when we examine relatively simple and straightforward measures that are commonly used to evaluate the fidelity of program implementation



(“programmatic content” for short), we found no statistically significant differences.

Specifically, Table 5 provides a comparison of these measures between the two treatment groups. Our analysis suggests that the quality of the government-organized teacher training for the CAL program did not differ from that of the NGO-organized teacher training. The teacher training evaluation scores of government schools were 1.25 questions (out of 20) lower than those of the NGO CAL program (statistically significant at the 5% level—see Row 1, Column 1). Given that almost all of the evaluation items were about setting up and operating the software, and constant tech support was provided for all government and NGO treatment schools throughout the duration of the intervention, 1.25 items is likely not a substantive difference. As for program length, the government CAL program was slightly (0.18 months) shorter than the NGO program, although the difference is not statistically significant (Row 1, Column 2). Similarly, the frequency with which CAL sessions were held in the two treatment groups did not significantly differ (either in magnitude or in a statistical sense—Row 1, Columns 3 and 4).

Table 5. Ordinary least squares estimates of the effects of implementers on training quality and duration/frequency of CAL program

		Dependent variable			
		Teacher training evaluation scores (0-20)	Number of months CAL was implemented	CAL sessions 2X per week-- reported by teacher (1=yes, 0=no)	CAL sessions 2X per week-- reported by student (1=yes, 0=no)
		(1)	(2)	(3)	(4)
[1]	Government schools (1 = yes, 0 = no)	-1.25** (0.49)	-0.18 (0.70)	0.13 (0.11)	-0.06 (0.10)
[2]	Control variables <sup>a</sup>	No	No	No	Yes
[3]	Observations	80	80	80	3,533
[4]	R-squared	0.08	0.00	0.02	0.08
[5]	Mean of NGO schools	12.57	5.33	0.28	0.34

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses.

<sup>a</sup> Control variables include the variables shown in Table 1.

Our data, however, show that the CAL program protocols were not fully followed by the administrators and teachers—especially in the government schools (Table 6). For example, administrators and teachers in 38 percent of the government schools replaced regular English classes with CAL classes in comparison to 18 percent of the NGO schools (a 20-percentage point difference which was statistically significant at the 5 percent level; Row 1, Column 1). In addition, instead of assigning computer teachers to run the CAL sessions as required by the program protocol, 48 percent of the government schools assigned English teachers while only 18 percent of NGO schools did the same. (This 30-percentage point difference was also statistically significant at the 1 percent level; Row 1, Column 2.) Based on these statistics, government schools were two to three times more likely than NGO schools to break these aspects of the protocol. While we do not have data to explain why these types of actions were taken, both indicators are consistent with an interpretation that the program protocol was broken in the government schools in a way to save resources (here, teacher effort). English teachers would spend less time preparing for and teaching their regular English classes, instead promoting students' self-study through the use of the CAL software (in place of the regular English classes). These students, therefore, ultimately received less teacher-led instruction than they did before the start of the CAL program.

Table 6. Ordinary least squares estimates of government versus NGO implementer effects on instructional substitution and direct monitoring

		Dependent variable		
		Regular CAL classes replaced English classes (1 = yes, 0 = no)	Appointed English teacher as the CAL supervisor (1 = yes, 0 = no)	Education bureau or NGO called or visited the school to follow up with program progress (1 = yes, 0 = no)
		(1)	(2)	(3)
[1]	Government schools (1=yes, 0=no)	0.20** (0.10)	0.30*** (0.10)	-0.50*** (0.08)
[2]	Observations	80	80	80
[3]	R-squared	0.05	0.10	0.33
[4]	Mean of NGO schools	0.18	0.18	1.00

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses.

The failure of the government CAL program to improve student achievement, coupled with the above-described substitution of instructional time, tentatively suggests that one reason the NGO CAL program succeeded (relative to both the government-run CAL schools and the control schools) was increased exposure to English. In other words, the schools in the NGO program had more English instruction, in the form of both CAL instruction and teacher-led instruction, than did the schools in the government CAL program and control group. Another possible interpretation of the findings is that since government schools had even less teacher-led instruction than control schools as a result of substitution, and neither government nor control schools affected English scores, that teacher-led instruction and CAL were equally effective in our study schools. If this is the case (and it is only speculative given the fact that the substitution of instructional time is not exogenous), it may not only be CAL, per se, that is driving the rise in outcomes in the NGO group, but both CAL and additional instruction time (or perhaps even only the additional instruction time).

Beyond the substitution of instructional time, we also examine treatment group differences in direct monitoring (Table 6, Column 3). Specifically, we find that government officials were much less likely than NGO staff to call or visit their program schools and follow up with program progress (50 percentage points, statistically significant at the 1 percent level). Although we are unable to definitively establish a causal link, the dearth of direct monitoring may have contributed to the program's failure (both in terms of the substitution of instructional time as well as the program's ability to raise test scores).<sup>5</sup>

To gain tentative insight into the relative importance of these different potential mechanisms (intermediate variables), especially the substitution of instructional time and direct monitoring, we examined correlations between the intermediate variables (from Tables 5 and 6) and student test scores (Table 7, Columns 1 to 5). The models in Table 7 all include basic controls. They then progressively add other different intermediate variables from Tables 5 and 6. The coefficients in Row 1 indicate a negative relationship between student test scores and the replacement of class time. The negative relationship holds across different specifications that control for other intermediate variables such as direct monitoring (Row 2), program frequency (Rows 3-4), the quality of teacher training (Row 5), program duration (Row 6), and the replacement of the computer teacher with the English teacher (Row 7). Importantly, the coefficient on replacement of class time is similar in magnitude and statistically significant when included alone (Column 1) or in combination with all other intermediate variables (Column 5). Furthermore, the coefficients in Row 2, Columns 2, 3 and 4 reveal a positive and statistically significant correlation between student scores and direct monitoring. The coefficient on direct

---

<sup>5</sup> There are of course other channels by which direct monitoring might affect the effectiveness of the program. A lack of direct monitoring might mean that there was little checking for whether students were using the software in the correct way (i.e., not working on remedial questions for the appropriate unit or not using the software for the total duration of class).

monitoring is still positive (although no longer statistically significant), even when holding constant intermediate variables that are likely ex post affected by direct monitoring (number of months that CAL was implemented as well as whether CAL sessions were held twice a week—reported by students or teachers—Rows 3,4, and 6). These correlational results suggest that the relative absence of direct monitoring in the government CAL schools may also help explain the poorer performance (in terms of standardized English test scores) of the students in those schools. While speculative, the results from Table 7 are consistent with a story that when monitoring was undertaken more intensively, administrators and teachers in the government-run schools were less likely to replace CAL classes and test scores improved. Of course, such conclusions are tentative, given the correlational nature of the findings.

Table 7. Correlations between student scores and intermediate (program implementation and monitoring) variables

	Dependent variable: Evaluation English score ( <i>SD</i> )				
	(1)	(2)	(3)	(4)	(5)
[1] Regular CAL classes replaced English classes (1 = yes, 0 = no)	-0.15** (0.07)		-0.11 (0.07)	-0.08 (0.07)	-0.14* (0.08)
[2] Education bureau or the NGO called or visited the school to follow up with program progress (1 = yes, 0 = no)		0.23*** (0.06)	0.21*** (0.06)	0.16** (0.07)	0.11 (0.08)
[3] CAL sessions 2X per week--reported by student (1 = yes, 0 = no)					-0.10 (0.07)
[4] CAL sessions 2X per week--reported by teacher (1 = yes, 0 = no)					-0.07 (0.08)
[5] Teacher training evaluation scores (0-20 pts)				0.02 (0.01)	0.02 (0.01)
[6] Number of months CAL was implemented					0.02 (0.01)
[7] Appointed English teacher as the CAL supervisor (1 = yes, 0 = no)				-0.09 (0.08)	-0.07 (0.08)
[8] Control variables <sup>a</sup>	Yes	Yes	Yes	Yes	Yes
[9] Observations	3,533	3,533	3,533	3,533	3,533
[10] R-squared	0.65	0.65	0.66	0.66	0.66

\* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

<sup>a</sup> Control variables include the variables shown in Table 1.

## 5 Cost-effectiveness analysis

We analyzed cost effectiveness of the government and NGO CAL programs using the method suggested by Dhaliwal et al. (2011). The program's main costs are teacher training and class subsidies; software design, development, and installation; and monitoring fees. From the perspective of policymakers who are considering upscaling the program, computer hardware is already a sunk cost, given that the government has been installing computer labs in every rural elementary school as part of its 13th Five-Year Plan. Altogether, the costs per unit of the NGO CAL program in terms of improvements in student learning are 42.8 USD per standard deviation (Appendix B).<sup>6</sup>

<sup>6</sup> In terms of its cost-effectiveness ratio of 42.8 USD/SD, the CAL program falls in the middle of the list of programs

Considering that government schools were supposed to be monitored by local government officials, but were frequently not, the only difference in project costs between the NGO-led and government-led project is the monitoring fee of 0.91 USD/student/year. All of the other costs for the two projects were identical. Overall, the additional expenditure for monitoring would induce a considerable impact on the learning of students.

## 6 Conclusion

This study showed that a CAL intervention was unsuccessful in raising student test scores when implemented by different institutions. We found that the CAL programs implemented in the two treatment arms (government and NGO) were identical in content, duration, frequency, and other important dimensions. However, compared to the NGO CAL program, the government CAL program was more likely to substitute out regular instruction and less likely to be directly monitored by implementers. Furthermore, substitution and direct monitoring were both correlated with changes in student test scores, tentatively suggesting that institutional context contributed to the disparity in student achievement between the two programs.

The study contributes to the limited literature on the role of institutions in program effectiveness in developing countries. Our results are consistent with the studies by Bold et al. (2013) and Banerjee et al. (2017), who found that implementers must engage in effective monitoring for interventions to be successful. Our results also suggest that educational programs run by NGOs may not yield the same results when run by governments. For instance,

---

examined in J-PAL (2014). It is lower than 7 programs (e.g. providing minimum conditional cash transfers in Malawi (1667.43), girls scholarships in Kenya (72.26), a Read-a-Thon program in the Philippines (85.07), and individually paced CAL in India (64.46)). It is higher than 6 programs (e.g. providing earnings information in Madagascar (0.85), electing and linking school committees to local governments in Indonesia (7.50), and remedial education in India (32.59)).

Muralidharan and Sundararaman (2015) showed that teachers in private schools are more effective and have lower absenteeism than teachers in government-run schools in India, despite lower pay, less knowledge/experience, and lower cost per student.

The tendency of teachers to take advantage of additional inputs and simultaneously reduce effort—behaviors that we observed in the government treatment arm—also finds parallels in the literature. Duflo et al. (2015) and Muralidharan and Sundararaman (2013) both find that providing additional resources to reduce the student-teacher ratio (such as contract teachers) can cause existing teachers to reduce their efforts at work. In keeping with these findings, Muralidharan et al. (2017b) estimated that increasing inspections and monitoring could be over ten times more cost effective at increasing teacher-student contact time (through reduced teacher absence) than hiring additional regular teachers. Hence, carefully conducted systematic monitoring should be in place when additional resources are given to the schools.

Our results may also help interpret studies that observe a lack of program effectiveness in complex, long-term education interventions. For example, Barrera-Osorio and Linden (2009) find that school-level programs in Colombia, which provide computers and teaching training, have no impact due to poor implementation. The program developers assumed that the teachers would cooperate and, therefore, did not engage in any monitoring. Further, the teachers did not incorporate the new technology into their teaching. The failure of this program to improve student outcomes, just like the failure of the program in the Bold et al. (2013) study, suggests that additional inputs and monitoring from implementing institutions may be required. In one-off programs that involve less complex interventions, such as the provision of information (e.g. Jensen 2010), scholarships (Kremer, Miguel and Thornton 2009) or subsidized health products



(e.g. Glewwe, Park, Zhao 2016; Dizon-Ross et al. 2017), the principal-agent problem may not be as severe.

Following our findings, future research could investigate how different types of institutions (e.g. public and private institutions) could cooperate to implement evidence-based interventions at scale, including how to improve the quality of monitoring and avoid substitution effects. Muralidharan and Niehaus (2017) suggest that researchers should create formal institutional frameworks for collaboration to avoid inefficiencies that arise from working with government partners. They suggest engaging in formal agreements, such as memorandums of understanding, and cite a number of successful interventions in which the NGO and government partners did so.

### **Conflict of Interest**

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

### **Acknowledgements**

The authors are supported by the 111 Project (Grant Number B16031), the National Natural Science Foundation of China (Grand Number 71333012, 71933003), and the International Initiative for Impact Evaluation (3ie).

## References

- Auriol, Emmanuelle, and Michael Walters. 2009. "The Marginal Cost of Public Funds and Tax Reform in Africa." *Journal of Development Economics* 97(1): 09-110.
- Bai, Yu, Di Mo, Linxiu Zhang, Matthew Boswell, and Scott Rozelle. 2016. "The Impact of Integrating ICT with Teaching: Evidence from a Randomized Controlled Trial in Rural Schools in China." *Computers & Education* 96: 1–14.
- Barrerra-Osorio, Felipe and Leigh L. Linden. 2009. "The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia." World Bank Policy Research Working Papers 4836. doi:10.1596/1813-9450-4836.
- Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31(4): 73-102.
- Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*. <http://www.nber.org/papers/w11904>.
- Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling up What Works: Experimental Evidence on External Validity in Kenyan Education." Center for Global Development Working Paper, no. 321. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2241240](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2241240).
- Bolton, Kingsley, and David Graddol. 2012. "English in China Today." *English Today* 28(3): 3.
- Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.
- 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.
- Dizon-Ross, Rebecca, Pascaline Dupas, and Jonathan Robinson. 2017. "Governance and the Effectiveness of Public Health Subsidies: Evidence from Ghana, Kenya and Uganda." *Journal of Public Economics* 156: 150-169.
- Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives, and Pupil-teacher Ratios: Experimental Evidence from Kenyan Primary Schools." *Journal of Public Economics* 123: 92-110.
- 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.
- Escueta, Maya, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos. 2017. "Education Technology: An Evidence-Based Review." NBER Working Paper No. 23744. JEL No. I20,I29,J24.
- Glewwe, P., Park, A., & Zhao, M. (2016). A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China. *Journal of Development Economics*, 122, 170-182.
- Hu, Guangwei. 2005. "English Language Education in China: Policies, Progress, and Problems." *Language Policy* 4 (1): 5–24. doi:10.1007/s10993-004-6561-7.

- Hu, X. 2009. "The Quality of English Teacher of the Primary School in Rural China-Evidence from Heilongjiang Province." *China Adult Education* 18: 89–90. [In Chinese]
- Jensen, R. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics* 125(2): 515–548.
- J-PAL. 2014. "Student learning and student attendance cost-effectiveness analysis data." Jameel Poverty Action Lab. [www.povertyactionlab.org/doc/cea-data-full-workbook](http://www.povertyactionlab.org/doc/cea-data-full-workbook) (accessed August 2014).
- Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3), 437–456.
- Lai, Chun, and Mingyue Gu. 2011. "Self-Regulated out-of-Class Language Learning with Technology." *Computer Assisted Language Learning* 24(4): 317–35. doi:10.1080/09588221.2011.568417.
- Li, Jongying. 2002. "An Empirical Study on Learning Strategies of Chinese ESL Learners." *Foreign Language Education* 1(1). [http://en.cnki.com.cn/Article\\_en/CJFDTOTAL-TEAC200201008.htm](http://en.cnki.com.cn/Article_en/CJFDTOTAL-TEAC200201008.htm).
- Loyalka, Prashant, Zhaolei Shi, James Chu, Natalie Johnson, Jianguo Wei, and Scott Rozelle. 2014. "Is the High School Admissions Process Fair? Explaining Inequalities in Elite High School Enrollments in Developing Countries." REAP Working Paper 276. California, USA: Freeman Spogli Institute of International Studies. [http://reap.fsi.stanford.edu/sites/default/files/276-is\\_the\\_high\\_school\\_admissions\\_process\\_fair.pdf](http://reap.fsi.stanford.edu/sites/default/files/276-is_the_high_school_admissions_process_fair.pdf).
- Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *The Quarterly Journal of Economics* 134(3): 1627–1673.
- McKay, Sandra Lee. 2002. *Teaching English As An International Language: Rethinking Goals and Perspectives*. New York: Oxford University Press. <http://www.tesl-ej.org/wordpress/issues/volume7/ej25/ej25r5/?wscr=>.
- Mo, Di, Weiming Huang, Yaojiang Shi, Linxiu Zhang, Matthew Boswell, and Scott Rozelle. 2015. "Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China." *China Economic Review* 36: 131–45.
- Ministry of Education (MOE). 2014. "China Education Yearbook 2014," Ministry of Education, China. Retrieved from: <http://video.moe.gov.cn/nianjian/zgjynj2014.zip>
- Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31(4): 103–24.
- Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *The Quarterly Journal of Economics* 130(3): 1011–1066. <https://doi.org/10.1093/qje/qjv013>.
- Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." NBER Working Paper 19440.
- Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian. 2017a. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." NBER Working Paper 22923.

- Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017b. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." *Journal of Public Economics* 145: 116-135.
- National Bureau of Statistics. 2014. "China Statistical Yearbook 2013."
- Pal, Joyojeet. 2006. "Early-Stage Practicalities of Implementing Computer Aided Education: Experience from India." In *Fourth IEEE International Workshop on Technology for Education in Developing Countries (TEDC'06)*, 26–30. IEEE.  
[http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=1648401](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1648401).
- Jerusalem, Matthias and Ralf Schwarzer. 1992. "Self-Efficacy as a Resource Factor in Stress Appraisal Processes." In R. Schwarzer (Ed.), *Self-Efficacy: Thought Control of Action*, 195–213. Washington, DC, US: Hemisphere Publishing Corp.
- Vivalt, Eva. Forthcoming. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association*. <http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf>.
- World Bank. 2003. "World Development Report 2004: Making Services Work for Poor People." <https://openknowledge.worldbank.org/handle/10986/5986>.
- Xi, Zhiqi, Yu Bai, Ai Yue, and Linxiu Zhang. 2017. Computer Usage in Northwestern Rural China. *New Silk Road* (4): 88-89. [In Chinese]
- Zhao, Hongyuan. 2003. "Exploring Resources of Local Courses: An Effective Way of Developing Teaching Body in Ethnic Minority Regions." *Journal of Research on Education for Ethnic Minorities* 14(4): 60–64.
- Zheng, H. 2009. "English Teaching and Learning in Rural China Primary School." *Journal of Teaching and Management*, 12, 82e83. [In Chinese]

## Appendix A. Calculating the impact of CAL on raw English scores

Table A1. Ordinary least squares estimates of the treatment effect on raw English test scores

	Dependent variable: Raw Evaluation English Score	
	(1)	(2)
[1] NGO schools (1 = yes, 0 = no)	2.13** (1.02)	2.13** (1.00)
[2] Government schools (1 = yes, 0 = no)	-0.90 (0.92)	-0.90 (0.91)
[3] Control variables <sup>a</sup>	No	Yes
[4] Observations	5,253	5,253
[5] R-squared	0.64	0.65

\*  $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ . Robust standard errors in parentheses clustered at school level.

<sup>a</sup> Control variables include the variables shown in Table 1.

## **Appendix B. Calculating the cost-effectiveness of the CAL program**

A number of factors are involved in the calculation of the cost of the CAL program. First, the cost of training teachers was 34.43 USD/teacher: this included the provision of hardcopy materials, transportation, and other miscellaneous resources. We also provided program teachers with a stipend of 163.93 USD every academic year. A program teacher taught 35 students on average. Therefore, the average cost was  $(34.43 + 163.93) / 35 = 5.67$  USD/student/year.

Second, the total cost of the CAL software was  $0.44 + 0.69 = 1.13$  USD/student/year. How was this calculated? The cost to design and develop the software was a one-time expenditure of 4,918 USD. Assuming that the software lasts for three years, its per-student unit cost is  $4,918 \text{ USD} / 3 \text{ years} / 3,751 \text{ students} = 0.44$  USD/student/year. Software installation costs  $24.1 \text{ USD/class} / (35 \text{ students/class}) = 0.69$  USD/ student/year.

Third, the monitoring fee was an average of 31.8 USD per class per year in NGO schools, including transportation, hotels, and stipends for on-the-ground program implementers. The total monitoring fee per student was  $(31.8 \text{ RMB/class/year}) / (35 \text{ students/class}) = 0.91$  USD/student/year.

We can approximate the public resource investment as 20 percent of the cost of program execution (Auriol and Walters 2009). Social costs include costs of the program's execution and public resource investment.

Table B1 below summarizes the above costs and presents the cost-effectiveness ratio for the NGO-led CAL program.

Table B1. Cost-effectiveness of the NGO CAL Program

	NGO-led (1)
I. Cost of program execution (USD/student/year)	
Training and stipends	5.67
Software development and installation	1.13
Monitoring	0.91
Subtotal	7.71
Public resource investment	1.54
Social cost	9.25
II. Effectiveness (in <i>SDs</i> )	
Program effect	0.18***
III. Cost-effectiveness ratio (USD/ <i>SD</i> )	
Program cost-effective ratio	42.8
Social cost-effective ratio	51.4

\*p < .10, \*\*p < .05, \*\*\*p < .01. P-values calculated using robust standard errors clustered at school level