1 Introduction

There has long been much interest in the possible existence of peer effects in learning. Examples in economics include natural experiments in roommate or squadron member assignment in institutions of higher education (Sacerdote 2001; Zimmerman, 2003; Lyle, 2007 & 2009; Carrell, Fullerton and West, 2009; Shue, 2012) and in the assignment of students to school through busing (Angrist and Lang, 2004) or random assignment to classes within schools (Vigdor and Nechyba, 2007; Graham, 2008). There is also more recent evidence from randomized experiments where the peer group composition is randomly varied by the researchers (Duflo, Dupas and Kremer, 2011; Carrel, Sacerdote and West, 2013).

It has long been suspected that peer effects in learning vary with the characteristics of the peers. The difficulty is in estimating this heterogeneity in a convincing manner. The existing studies on peer effects vary largely in estimates (Sacerdote, 2011). Part of the reason may be that many studies adopt a Linear-in-Means model that assumes that peer effects are homogeneous across students (Hoxby and Weingarth, 2005). Ignoring the heterogeneity of peer effects among students with different characteristics can lead to misleading conclusions about the existence or the magnitude of peer effects. Furthermore, using concurrent outcomes of the peer group to identify peer effects on own outcome cannot distinguish real peer effects from common shocks that affect the whole group (Sacerdote, 2001). It can also be difficult to identify the true peer group with whom a student interacts. For instance, increasing the number of high-achieving students in a group may induce low-achieving students to form subgroups among each other, something they might not have done if high-achieving students are less available (Carrel, Sacerdote and West, 2013). Introducing exogenous changes in peer groups often is crucial to obtaining correctly estimated peer effects.

A major challenge in properly identifying heterogenous peer effects is to suitably control for all possible effects of treatment that are not driven by peer effects themselves. In this paper we tackle this challenge using data from a large-scale randomized controlled trial that allocates primary school students to computer assisted learning (CAL). We take advantage of the fact that randomization takes place at three levels: (1) assignment of schools to CAL treatment and control; (2) assignment of students to CAL treatment either individually or in pairs; and (3)

random assignment of a peer for those students assigned to treatment in pairs. We also have baseline data on all students, including results from a standardized academic test. We show that these different pieces of information are needed to identify the heterogeneous effect of receiving the CAL treatment in pairs. Although it is possible to draw inference about how the effect of treatment varies by the type of peer with less data, it is not possible to establish the sign of the effect itself – and thus it is not possible to draw policy conclusions – without suitable control groups. This problem is common to studies in which all subjects are paired, including several of the studies cited above.

Our results indicate that the average effect of computer assisted learning is the same whether student receives the treatment individually or in pairs: on average, students do not learn more (or less) if they receive CAL individually. This has important budgetary implications since it is half as expensive to treat students in pairs compared to individual treatment. We also find significant heterogeneous peer effects. Weaker students benefit more from CAL if they are paired with a stronger student, while stronger students learn more when they are paired with a weaker student. In contrast, students of average ability benefit equally from CAL treatment irrespective of the initial ability of the student they are paired with.

These findings contribute to the existing literature in several ways. The study adds to the general understanding of peer effects estimated from experimental evidence. In particular, we believe this is the first study that estimates peer effects by randomly pairing students for a specific learning activity in class. Our study highlights the importance of heterogeneous peer effects in this context. The evidence provided in our study suggests that learning can be enhanced by optimally pairing students for joint learning activities – in our case, by pairing low and high-achieving students together. We suspect that this arises because strong students get an even better understanding of the material when they try to explain it to their weaker peers. These conclusions about pairing have important policy implications for a cost-effective delivery of computer assisted learning programs in China and elsewhere. Such findings complement ongoing work estimating the average treatment effects of CAL in different regions of China and among different rural populations (e.g., Lai et al., 2011, 2012 and 2013; Mo et al., 2013 and 2014).

The paper is organized as follows. The experimental design is summarized in Section 2. In

Section 3 we present our testing strategy in detail, contrasting what can – and cannot – be inferred with different types of data. The student data are described in Section 4. Estimation results are detailed in Section 5.

2 The experiment

During the 2011-2012 academic year, we conducted a randomized controlled trial to study peer effects in Computer Assisted Learning (or CAL) in China. The main focus of the CAL intervention is remedial tutoring in mathematics to complement the regular school curriculum. CAL is not intended to help top student performers advance faster and learn more than the school curriculum. It aims instead at helping weaker students keep up with the rest of the class. What is unclear is whether it is capable of reaching this objective.

2.1 Experimental design

One of the objectives of the study is to identify interventions that can bridge the educational gap between rich and poor Chinese counties. For this reason, we implement the randomized controlled trial in a poor area of China. We select the Shaanxi Province, a province with one of the greatest number of nationally designated poor counties (CNBS, 2013). Within Shaanxi, we choose to focus on the Ankang prefecture because it is the poorest prefecture in the province (CNBS, 2013). Of the eight counties in Ankang that are nationally-designated as poor (CNBS, 2013), we randomly select four. With an average per capita income of 4000 RMB ($650) per year in 2011, the four selected counties have an average income that is far below the rural China average, which was 6977 RMB in 2011 (CNBS, 2011). All 72 six-year primary schools in the four selected counties are included in the experiment. Within sample schools, we work with students in grades three to six because the CAL software was produced for these grades.[1] All classes in these grades are included. None of our sample students had ever participated in a CAL program prior to the 2011-12 academic year. A total of 7881 grade students were involved in the study.

Half of the 72 sample schools were randomly assigned to receive the CAL treatment and the other half were assigned to the control group. When dividing schools into treatment and control,

[1] Grade 1 and 2 students are not included because they can not read at levels high enough to use software.

we pre-balanced on the student and family characteristics reported in Table 2, following the methodology suggested by Bruhn and McKenzie (2009). The CAL intervention was implemented during over the entire 2011-2012 academic year. Students in treated schools received two 40-minute CAL sessions per week. The sessions took place in the school and they were mandatory for all students in treated schools.

Protocols are designed to ensure that the control schools provide a true counterfactual. Students in the 36 control schools took their regular math classes as usual, without any CAL intervention. To avoid spillover effects across schools, the principal, teachers, students, and parents in the control schools were not informed of the CAL project. The research team did not visit the control schools except for the baseline and endline surveys. No placebo activity was organized in control schools. Treatment thus represents additional teaching time. The possibility of accidental spillover is minimized by the fact that there was only one sample school per town. This means that the average distance between control and treatment schools is more than 30 kilometers. No student in a treatment school lived in a village with a student from a control school.

During CAL sessions, students played math games designed to help them review and practice the material taught during their regular math classes. The instructional videos and games that make up the content of the CAL software are all based on the material in text books that use rural China's most common curriculum, the uniform national curriculum. The content is grade-specific and is the same across all treated schools for students in the same grade. In a typical session, the students first watch an animated video that reviews the material taught by their math teacher during that week. The students then play games containing various math exercises. The games have cartoon characters and story lines that make the exercises fun.

Many CAL students were randomly assigned a peer who was a student in the same class. The pair shared a computer during the CAL sessions. Peers were assigned randomly by the research team from among the students in the same class. Peer assignment was decided once and for all at the beginning of the academic year after the baseline survey, and it remained unchanged over the duration of treatment. To keep a log of which students shared a computer in each CAL session, students were required to log in using their unique username and password. According to log records, there was almost no switching of peers across sessions. Furthermore, less than

one percent of paired students participated in a session alone due to their peer's absence on the day of the session.

Because some classes have an odd number of students, six percent of students in treated schools were not assigned a peer. As as result, the sample includes both paired and unpaired students. Unpaired students participated in the same CAL sessions as the paired students, but they did not have to share their computer with anyone. In addition, some students lost their peer when the peer left the school in the middle of the school year. These students were not reassigned a peer. If a student participated to more than half of the CAL sessions without sharing a computer, this student is categorized as unpaired for the purpose of our analysis. This only affects 23 students in 7 schools.

The experimental protocol was designed to minimize interaction with students other than one's peer. During CAL sessions, paired students were allowed to interact freely, but no discussion or interaction was allowed with other students. Sharing a pair of earphones also helped paired students focus their attention and conversations on their own computer, and limit conversations with others. The teacher who supervised a CAL session was only allowed to help students with scheduling, computer hardware issues, and software operation. The main duty of the teacher-supervisor was to ensure that each weekly CAL sessions matched the pace of regular math classes. According to our own in-class observations, the sessions were so intense that the students had little time or interest to interact with other student pairs or with the teacher-supervisor.

2.2 Data collection

We conducted two survey rounds in the 72 sample schools – one at baseline in June 2011 and one at endline in June 2012. All students in grades three to six participated in the two rounds of surveys. The baseline survey was conducted at the end of the spring semester, before any implementation of the CAL intervention had begun. The endline survey was conducted in June 2012 after the intervention had been running for an entire academic year. The two survey rounds are almost identical in terms of design and questionnaire. Information includes the gender of the student, whether the student is an only child, whether the student had prior computer experience, and whether the student's mother and father are illiterate.

During each survey round, the enumeration team visited each school and gave all students a standardized math test. The test is a grade-specific multiple choice test and is identical for all students in the same grade. The questions are all chosen from the TIMSS test data bank. Elementary teachers in rural schools of Shaanxi Province screened the questions to ensure that they were appropriate, i.e., neither too difficult nor too easy for the average student. None of the questions repeat questions used as exercises in the CAL software. The test takes 25 minutes and was administered using pen and paper so as not to advantage CAL students. Since students take a grade-specific test, scores are not directly comparable across baseline and endline. To make test scores comparable, they have been standardized using grade-specific test scores obtained by control students. Throughout the analysis, math scores are measured in terms of standard deviation units relative to the average score of control students.

3 Testing strategy

Our aim is to obtain consistent estimates of the heterogeneous effect of CAL treatment on paired students. In this section we discuss how this can be achieved using the data at our disposal. The pros and cons of different estimation approaches are briefly discussed before we settled on our preferred estimation strategy. Discussing different possible methodologies in some detail will save much time when we present the results themselves.

We need to distinguish between three types of treatment effects: (a) the average treatment effect of CAL; (b) the average treatment effect of taking CAL in pairs rather than individually; and (c) the effect of having been assigned a particular peer, conditional on being paired. The first effect (a) is the focus of earlier work by Lai et al. (2011, 2012 & 2013) and Mo et al. (2013 & 2014). These studies estimate the average treatment effects of CAL in different regions of China and among different rural populations. The estimated program impacts range from 0.12 standard deviations of a one-semester program among migrant students to 0.26 standard deviations of a three-semester program among rural students. The third effect (c) is what we focus on here. The question is whether we can obtain a consistent estimate of (c) without also consistently estimating (a) and (b).

3.1 Unpaired students

Control students measure the average performance of children without CAL treatment, while unpaired CAL students measure the average performance of CAL without peer effects. For unpaired students, the effect of CAL treatment can be written:

$$y_{it+1} = y_{it} + h(y_{it}) + f(y_{it})T_t + u_{it+1} \tag{1}$$

where y_{it} denotes the performance of student i in the math test at time t, $T_i = \{0, 1\}$ is a dummy for being assigned to CAL treatment, and $P_i = \{0, 1\}$ is a dummy for receiving the CAL treatment in pairs.

In model (1) $h(.)$ denotes the learning that takes place without treatment. This is estimated from the control population, and in general it varies with the initial level of the student y_{it}. For instance, if a student has already learned a topic, further instruction in that topic will not improve his/her knowledge of that subject. We expect $h(.)$ to be positive on average because students above the mean at baseline have a higher likelihood of being above the average at endline, except for regression to the mean due to measurement error or random performance variation on the test. The yet-to-be-defined function $f(.)$ captures the heterogeneous effect of treatment conditional on initial knowledge. For instance, if treatment has a stronger effect on initially weak students, then $f(.)$ is an decreasing function.

With a sufficiently large number of observations, we could in principle estimate a flexible version of model (1). Unfortunately we do not have that luxury. A linear version of model (1) is of the form:

$$y_{it+1} = k + \rho y_{it} + (\alpha + \gamma(y_{it} - \overline{y}_t))T_t + u_{it+1} \tag{2}$$

where we have explicitly demeaned y_{it} in the interaction term so that α can be interpreted as the average treatment effect (Wooldridge, 2003).

The intercept k is the average unconditional level of knowledge at $t + 1$ without treatment, $\rho - 1$ is the average growth rate in knowledge, α is the average effect of the CAL treatment on all students, and γ is the heterogeneous effect of treatment depending on initial knowledge. If the treatment helps weaker students catch up, then $\gamma < 0$: initially knowledgeable students

benefit less from treatment.

3.2 Paired students

For paired students, the total effect of treatment can be written as:

$$y_{it+1} = y_{it} + h(y_{it}) + f(y_{it})T_t + g(y_{jt}|y_{it})P_t + u_{it+1} \tag{3}$$

where $g(.)$ is an unknown function that captures peer effects. By experimental design $T_t = 1$ whenever $P_t = 1$ – i.e., only students who take CAL are paired. In our estimation, we posit $g(.)$ to be of the form:

$$g(y_{jt}|y_{it}) = \beta_0 + \beta_1(y_{it} - \overline{y}_t) + \beta_2(y_{jt} - \overline{y}_t) + \beta_3(y_{it} - \overline{y}_t)(y_{jt} - \overline{y}_t) \tag{4}$$

where we have demeaned all y's to facilitate interpretation of the parameters. The interpretation of each coefficient is as follows: $\beta_0 > 0$ is the average incremental gain in learning for a student of average initial knowledge paired with an average peer, compared to an unpaired student of similar ability; $\beta_1 < 0$ means that a student i with high initial knowledge benefits from CAL less if paired than if not paired; $\beta_2 > 0$ means that a student i benefits more from CAL if paired to a student j with high initial knowledge than if paired with an average peer; and $\beta_3 < 0$ means that a student i of high initial knowledge benefit less from CAL if paired with another high knowledge student j compared to being paired with an average peer. More formally, we have:

$$\begin{aligned} \frac{\partial g}{\partial y_{it}} &= \beta_1 + \beta_3(y_{jt} - \overline{y}_t) \\ \frac{\partial g}{\partial y_{jt}} &= \beta_2 + \beta_3(y_{it} - \overline{y}_t) \\ \frac{\partial^2 g}{\partial y_{it}\partial y_{jt}} &= \beta_3 \end{aligned}$$

Combining (2) with (4) the estimated model is:

$$\begin{aligned} y_{it+1} &= k + \rho y_{it} + (\alpha + \gamma(y_{it} - \overline{y}_t))T_t \\ &\quad +(\beta_0 + \beta_1(y_{it} - \overline{y}_t) + \beta_2(y_{jt} - \overline{y}_t) + \beta_3(y_{it} - \overline{y}_t)(y_{jt} - \overline{y}_t))P_t + u_{it+1} \end{aligned} \tag{5}$$

Coefficient α measures (a), the average treatment effect of CAL and coefficient β_0 measures (b), the average treatment effect of being paired for treatment. Peer effects (c) are captured by coefficients β_1, β_2 and β_3.[2]

3.3 Class effects

So far we have assumed that CAL and pairing have an effect that depends on the absolute level of initial knowledge of students and their peers. It is also possible that what matters is the initial knowledge of a student relative to others in the class. This could arise, for instance, if teachers teach to the class, i.e., go through the curriculum faster or deeper if the average student is stronger/is learning faster. In this case, CAL may help laggard students to catch up.[3]

To capture this possibility, we include $\overline{y}_{ct}$, the average initial knowledge of the class, as additional regressor, and we enter all interaction terms as deviation to the class mean $\overline{y}_{ct}$.[4] The estimated model becomes:

$$\begin{aligned} y_{it+1} &= k + \rho_0 y_{it} + \rho_1 \overline{y}_{ct} + (\alpha + \gamma(y_{it} - \overline{y}_{ct}))T_t \\ &\quad + (\beta_0 + \beta_1(y_{it} - \overline{y}_{ct}) + \beta_2(y_{jt} - \overline{y}_{ct}) + \beta_3(y_{it} - \overline{y}_{ct})(y_{jt} - \overline{y}_{ct}))P_t + u_{it+1} \end{aligned} \tag{6}$$

Estimating this model is the focus of the empirical part of the paper.

3.4 The golfer model

It is useful to compare our preferred model (6) to an alternative model used by Guryan, Kroft, and Notowidigdo (2009) to estimate peer effects among golfers. Indeed there are many similarities between their experimental design and ours, given that golfers are randomly assigned to play in pairs. Guryan et al. wish to estimate whether a golfer plays better if paired with

[2]The β coefficients should be understood as capturing both exogenous and endogenous peer effects (Manski 1993), i.e., the effect of being paired with a treated student j, and the multiplier effect of j's CAL-induced learning on i's own learning. To estimate endogenous and exogenous effects separately, we would either need to observe paired students who did not to receive CAL treatment, or observe students paired with different numbers of peers (e.g., Fafchamps and Vicente 2014; Fafchamps, Vaz and Vicente 2014). Neither of these is possible here given the design of our intervention.

[3]Even if relative performance does not matter, we still may want to include average class performance as regressor to control for class differences that may, in a small sample, be correlated with treatment.

[4]The reader may wonder whether, in model (6), α can still be interpreted as the ATE of the CAL intervention even though we have not subtracted the mean of $(y_{it} - \overline{y}_{ct})$ from each interaction term. The answer is yes because the mean of $(y_{it} - \overline{y}_{ct})$ is, by construction, equal to 0.

a good golfer than if paired with a bad golfer. Let y_{it+1} be the performance of golfer i in the tournament, and let y_{jt} be the past performance of the paired player. The model that Guryan et al. estimate is of the form:

$$y_{it+1} = \beta_0 + \beta_2 y_{jt} + u_{it+1} \tag{7}$$

only using data on paired subjects, i.e., with $P_i = 1$. This is because, by design, tournaments only include paired golfers.

In model (7) there is an exclusion bias because y_{it} is positively correlated with y_{it+1} – and hence with e_{it+1} – but negatively correlated with y_{jt}. This negative correlation arises mechanically because good golfers are, on average, paired with golfers that are worse than them, while bad golfers are, on average, paired with golfers that are better than them. The solution Guryan et al. propose is to add y_{it} as control to eliminate the bias:

$$y_{it+1} = \beta_0 + \beta_1 y_{it} + \beta_2 y_{jt} + u_{it+1} \tag{8}$$

Caeyers (2013) shows that exclusion bias is largest when the size of the randomly assigned peer group is small – e.g., a pair.

Our model (5) can be seen as an extension of (8) to allow β to depend on the initial ability of golfer i. If, as in Guryan et al., we limit the estimation sample to paired subjects only, model (5) can be rewritten as:[5]

$$\begin{aligned} y_{it+1} &= (k + \beta_0 + (\rho + \gamma)\overline{y}_t) + (\rho + \gamma + \beta_1)(y_{it} - \overline{y}_t) \\ &\quad + \beta_2(y_{jt} - \overline{y}_t) + \beta_3(y_{it} - \overline{y}_t)(y_{jt} - \overline{y}_t) + u_{it+1} \end{aligned} \tag{9}$$

[5]This is obtained by using:

$$\begin{aligned} y_{it+1} &= k + \rho y_{it} + \gamma y_{it} T_{it} \\ &\quad + (\beta_0 + \beta_1(y_{it} - \overline{y}_t) + \beta_2(y_{jt} - \overline{y}_t) + \beta_3(y_{it} - \overline{y}_t)(y_{jt} - \overline{y}_t))P_t + u_{it+1} \\ &= (k + \beta_0 - \beta_1\overline{y}_t) + (\rho + \gamma + \beta_1)y_{it} + \beta_2(y_{jt} - \overline{y}_t) + \beta_3(y_{it} - \overline{y}_t)(y_{jt} - \overline{y}_t) + u_{it+1} \end{aligned}$$

from which we get our model in Guryan form:

$$\begin{aligned} y_{it+1} &= (k + \beta_0 - (\beta_1 + \beta_2)\overline{y}_t + \beta_3\overline{y}_t^2) + (\rho + \gamma + \beta_1 - \beta_3\overline{y}_t)y_{it} \\ &\quad + (\beta_2 - \beta_3\overline{y}_t)y_{jt} + \beta_3 y_{it} y_{jt} + u_{it+1} \end{aligned}$$

where we preserved the original notation.

The above shows two things. First, since we include past performance as regressor, our estimation model corrects for exclusion bias. Secondly, if we only use observations on paired students we cannot estimate ρ, γ and β_1 separately from each other. In other words, we cannot distinguish whether better able students perform better when paired (β_1), from whether students perform play better with CAL (γ), and from whether students who did well at baseline also perform better at endline (ρ). We can, however, still obtain consistent estimates β_2 and β_3. But without an estimate of β_1 we cannot compute the correct marginal effects of treatment. We will illustrate this in the empirical section.

Model (9) can be modified to include class effects as in (6). The same observation holds: since the mean of $(y_{it} - \overline{y}_{ct})$ is always 0 by construction, the interpretation of the coefficients is the same as above.

4 The data

A total of 7881 students from 72 primary schools were involved in the study. This total can be broken down into 1555 grade three students, 1927 grade four students, 2115 grade five students, and 2284 grade six students (Figure 1). There are 3852 students in the CAL schools and 4029 students in the control schools. Ninety-six percent of the students (3679) in the CAL schools have a peer with whom they shared a computer during the CAL sessions. The rest, i.e., 173 students sat alone without sharing a computer. As stated above, unpaired students arise mostly in classes with an odd number of students. On average, there was one student who had no peer in every two classes.

Table 1 presents information about balance across the three different types of treatments implemented in our experiment. We compute balance with respect to performance on the June 2011 math test and for the student characteristics collected in the baseline survey. The first two columns of Table 1 report regression coefficients of the variables listed on the left on treatment dummies. The comparison is between treated and control students and the dummy is 1 in treated schools and 0 in control schools. Results show that random assignment of CAL treatment across schools produced balanced groups of students in the CAL and control schools along all available

variables.

The next two columns of Table 1 compare paired and unpaired students. Here the comparison is between students who are treated individually and those who are treated in pairs. The dummy is 1 for those treated in pairs, and 0 for those treated individually. We do not find any significant difference between the two groups in terms of baseline characteristics. From this we conclude that randomization was successful and balanced is achieved on baseline characteristics.

The last two columns check random peer assignment for those treated in pairs. This is important given our emphasis on estimating heterogenous peer effects: if, in spite of our best efforts, peers are not assigned randomly, we worry that paired students may have been matched on unobservables, a feature that may introduce correlated unobservable effects and contaminate our inference.

In column (5) the Table reports regression coefficient of the baseline characteristic of one student on the baseline characteristic of the other. The estimated regression is of the form:

$$y_{it} = \beta_0 + \beta_2 y_{jt} + u_{it} \tag{10}$$

As discussed earlier, such random assignment test is subject to exclusion bias: because a student cannot be his/her own peer, negative correlation between peer characteristics naturally arises under random assignment. Consequently, under the null hypothesis of random assignment estimated $\widehat{\beta}_2$ are not centered on 0 but on a negative number. Caeyers (2013) derives the magnitude of the bias for groups and selection pools of fixed size and shows that the bias is particularly large when the randomly assigned group is small, e.g., in pairs. We cannot use Caeyers formula here because the size of the selection pools varies: class sizes are not constant. To circumvent this problem, we simulate the distribution of $\widehat{\beta}_2$ under the null using a so-called permutation method. This method also delivers a consistent p-value for β_2 and thus offers a way of testing the null of random assignment. This method works as follows. The object is to calculate the distribution of $\widehat{\beta}_2$ under the null that y_{it} and y_{jt} are uncorrelated. To simulate $\widehat{\beta}_2$ under the null, we create counterfactual random matches and estimate (10). In practice, this is implemented by artificially scambling the order of students within each class to reassign them into counterfactual random pairs. By construction these samples of paired observations satisfy the null of random

assignment within classroom. We repeat this process 1000 times to obtain a close approximation of the distribution of $\widehat{\beta}_2$ under the null. We then compare the actual $\widehat{\beta}_2$ to this distribution to get its p-value.

We present in Figure 2 the simulated distribution of $\widehat{\beta}_2$ for baseline math scores under the null hypothesis of random assignment. These simulated $\widehat{\beta}_2$'s are centered around -0.05, with very few values at or above 0. As shown in the first line of column (5) in Table 1, the $\widehat{\beta}_2$ estimated from the sample -0.03. Comparing this number to the histogram of $\widehat{\beta}_2$ under the null reported in Figure 2, we find that 27% of simulated coefficients are larger than -0.03. From this we conclude that the p-value is 0.27: we cannot reject the null hypothesis of random assignment based on baseline math scores.

In column (5) and (6) of Table 1 we report the coefficient estimates for other baseline characteristics as well as similarly calculated p-values for the null hypothesis of random assignment by these characteristics. All p-values are above the 10% level. From this we conclude that the random assignment of peers was implemented in a satisfactory manner.

Attrition during the experiment is low. A total of 7536 sample students surveyed in the baseline participated in the endline survey. Only 4% of the students who took the baseline survey did not take the endine survey. Based on information provided by the schools, attrition is mainly due to illness, dropout, and transfers to schools outside of the town. In Table 2 we examine whether attrition is correlated with treatment. Column 1 shows that attrition rates do not differ statistically between CAL school students and control school students. Attrition is also not correlated with being paired or not (Table 2, column 2) or with being assigned to a high or low achieving peer Table 2, column 3).

As a final check, we repeat the balancedness tests of Table 1 using only the non-attriting sample. The same conclusions hold: we cannot reject balance on all baseline characteristics for the first two treatments. We also repeat the permutation tests to check random peer assignment on baseline math scores. We obtain p-values all above 0.1 and again fail to reject the random peer assignment hypothesis.

5 Empirical analysis

In the first column of Table 4 we report coefficient estimates for model (9), the 'golfer' model in which we only use data on paired students. The mean math score of the class at baseline $\overline{y}_{ct}$ is included as control. The other estimate coefficient are shown interacted with P_i since, by construction, only paired students are used in the regression. As explained in Section 2, coefficient [6] estimates $\rho + \gamma + \beta_1$, the combined effect of past performance on its own ρ, interacted with CAL treatment γ, and interacted with being paired β_1. This coefficient is statistically significant, but we do not know which of the three effects it captures. Coefficient [8] is an estimate of β_2 while coefficient [10] is an estimate of β_3. We note that β_3 is significant and negative, which suggests that a low ability student benefits more from CAL if paired with a high ability student – or vice versa. Without an estimate of β_1 we cannot compute $g(.)$ in (4) and thus we cannot tell whether the absolute effect of CAL treatment is higher for high or low ability students.

By using data on control and unpaired students, we are able to separately estimate ρ, γ and β_1. This is done in the third column of Table 4, which estimates model (6) on the entire population of non-attriting students. Coefficient [1] is an estimate of ρ, which measures the extent to which performance in the June 2011 math test helps predict performance in the June 2012 math test. Since $\rho < 1$, this indicates math test scores exhibit a strong element of regression to the mean. This might arise because math test scores are noisy measures of math ability. Another possibility is that it signals convergence towards an average level of math proficiency. Since the purpose of our experiment is not to distinguish between the two, we do not pursue this issue any further. Coefficient [3] is an estimate of the average treatment effect of the CAL intervention, which is positive, statistically significant, and large in magnitude. This estimate is discussed in detail in Mo et al. (2014).

More of interest here is coefficient [4], which is an estimate of γ. This coefficient is indistinguishable from 0, indicating that the average positive effect of CAL on math performance is the same across students, irrespective of past performance. If this coefficient had been negative, we would have concluded that CAL helped laggard students catch up with their better performing peers. This is not what we find. A zero γ implies that, by itself, CAL is unable to reduce the

performance gap between students in a class. We observe a similar finding regarding β_1, which corresponds to coefficient [6] in column 3: the coefficient is slightly positive, but nowhere near statistically significant. In other words, students who did poorly on the June 2011 math test did not benefit more from CAL when paired than students who did well on that test. Taken together, these findings indicate that coefficient [6] in column 1 is entirely driven by ρ, that is, by coefficient [1] in column 3. This is exactly what we find: the coefficients are identical in magnitude and in significance.

Using coefficient estimates from column 3, we report in Table 5 the predicted performance of paired students at the June 2012 math test. Predictions are calculated for various hypothetical pairings of students with different levels of initial ability. The first row of the Table reports the predicted June 2012 performance of students who did quite poorly on the June 2011 test, that is, who received mark that is two standard deviation below the average. The first column is the predicted performance of such a student if he/she were paired with a student who did equally poorly on the June 2011 test. This predicted performance is -0.95, that is, just shy of one standard deviation below the average June 2012 test score. As emphasized earlier, there is random variation in test results for the same student over time, and thus considerable regression to the mean: someone who did exceptionally poorly in June 2011 must have had an unusually bad day, and their performance is predicted to improve in June 2012.

Moving to the other columns of row 1, we see that the predicted performance of an unusually poorly performing student improves if this student is paired with a better performing student during the CAL intervention: if such a student were paired with a top performer in 2011, their predicted performance would rise to -0.63, that is, 0.63 standard deviations below the 2012 test score average. We test whether the difference between columns 1 (-0.95) and 5 (-0.63) is statistically significant and we report the p-value of this test in the last column of Table 5. We find that the difference is significant at the 2% level, implying that a poorly performing student benefits more from CAL if paired with a high performer. A statistically significant effect of being paired with a good performer is also found in the second row of Table 5, that is, for students who received a score one standard deviation below average in June 2011.

In contrast, for a student who received an average score in 2011, we find no statistically significant relationship between predicted performance and the performance of the paired student.

In other words, the predicted performance of an average student is the same irrespective of the past performance of the student they are paired with during the CAL treatment. A similar result is found for students who received a mark one standard deviation above the average in the June 2011. For students who performed exceptionally well in 2011, we find that their predicted 2012 performance is, if anything, higher if they were paired with a poorly performing student: +1.21 compared to +0.99 standard deviation above the mean. This difference, however, is not statistically significant at conventional levels (p-value of 14%).

To test the robustness of our findings to alternative functional form assumptions, we reestimate models (9) and (6) with additional quadratic terms (coefficients [7] and [9]). Results are shown in columns 2 and 4 of Table 4, respectively. We find some evidence of non-linearity for paired students with respect to own 2011 scores. Other coefficients are largely unaffected. We report in Table 6 the performance predictions obtained using coefficient estimates reported in column 4 of Table 4. These calculations confirm the findings from Table 5. Students who performed one or two standard deviation below average in 2011 do better in 2012 if they are paired with high performers (significant at the 6% and 8% level, respectively). In contrast, high performers in 2011 do not do less well in 2012 if paired with poor performers; this difference is large in magnitude, albeit not statistically significant.

Tables 5 and 6 demonstrate that treatment effects vary across pairings. In Table 7 we present, for each of the pairings in Table 5, the predicted effect of CAL treatment relative to control students. The Table also reports pairing-specific p-values for the significance of the effect relative to controls. What the Table shows is that significant benefits from CAL are concentrated on two groups: (1) average and below-average students paired with above average-students; and (2) above-average students paired with below average students. The first group corresponds to the last two columns of the first three rows, where the estimated treatment effects of paired CAL are all positive and statistically significant at the 10% or better. The second group corresponds to the last two rows in columns one and two, with p-values less than 0.1. For weak students paired with weak students, the point estimate of the ATE is negative (row 1, column 1), although it is not statistically significant.

5.1 Improved pairing

Table 7 has shown that peer effects are stronger for some pairings than others. This suggests that it may be possible to increase the average treatment effect of CAL on math scores by favoring negative assorting, that is, by pairing weak students with strong students. To investigate the magnitude of this potential effect, we hypothetically match the weakest students with the strongest students in each class and calculate the predicted effect of CAL using the coefficients estimated in Table 4 (column 3).[6]

To implement this idea, we proceed as follows. We begin by sorting all the students in a class by their 2011 math score. We then pair the first student from the top with the first from the bottom, then the second from the top with the second from the bottom, and so on until every student is paired (if the number of students in the class is even) or until the median student is left to be treated individually (if the number of students in the class is odd). We can then compute the predicted treatment effect for each individual in the sample conditional on this improved match. Finally we aggregate these predicted effects to obtain the average predicted effect of the optimal match.

To recall, in the data the average treatment effect of CAL is a 0.17 SD improvement in math score. Based on our calculations, this improved pairing would further improve math test scores of paired students by another 0.03 SD relative to random pairing. This is equivalent to an 18% increase in treatment effectiveness on average. The difference between improved and random pairing is even larger – 0.04 SD – for weaker students, that is, for those with a 2011 math score below the class average. Improved pairing could thus be particularly beneficial to weak students.

5.2 Dispersion in math scores

We have seen from Tables 5 to 7 that students at both extremes of the score distribution gain more from CAL, especially if they are optimally matched. By itself, however, this does not tell us whether CAL leads to a reduction or an increase in the dispersion of math scores in treated classes. In other words, it does not tell us whether the improvement in math scores is achieved

[6]We do not claim that such pairing is optimal. Finding the pairing that maximizes average gains would probably require calculating, for each class, the value of predicted endline scores for each possible pairing of students in the class. While this is not impossible to implement in theory, negative assorting is much easier to implement in practice and is thus a more realistic policy. Booij, Leuven and Oosterbeek (2014) discusses a variety of assignment rules in the context of the assignment of university students to tutorial groups.

by helping weak students to catch up or by helping strong students to get further ahead of their peers.

To investigate this important issue from a policy point of view, we first note that the average improvement in math scores is 0.16 SD for students who scored higher than or equal to the class median in 2011. In contrast, the average improvement in scores is 0.19 SD for the students who scored lower than the class median in 2011. We further note that 9% of the average treatment effect of 0.17 is attributable to the "catching up" of the poorer performing students. From this we suspect that CAL reduces the dispersion in math scores for paired students compared to controls.

We can also look at the dispersion in scores directly. To this effect, we present in Table 8 various interdecile ranges for control and paired students. The first row reports the difference in standardized math scores between the 90th percentile (Q9) and the 10th percentile (Q1) students. This difference is 2.67 standard deviations for control students and 2.61 for paired students. Similar findings are shown in row 2 – which compares the 80th to the 20th percentiles – and in row 3 – which compares the 70th to the 30th percentiles. These results suggest that CAL reduced the dispersion in math scores among the treated population. In other words, students who were initially weak benefitted more than students who were initially strong.

Because interdecile differences are small in magnitude, we wonder whether they are statistically significant. To obtain a p-value for each of the three columns of Table 8, we use a method that has the advantage of being entirely non-parametric. Our null hypothesis is that the distribution of scores among the control and treatment populations is the same. We want to compare each of the interdecile differences in Table 8 to the distribution of interdecile differences that would arise under the null. To derive the distribution of these differences under the null, we simulate it from the data by randomly drawing hypothetical controls and treatments from the pooled observations, keeping the number of controls and treated identical to the actual data. In practice, this is achieved by randomly re-sorting the pooled data and assigning the first N^c observations to controls and the others to treated – where N^c is the number of control observations in the actual data.[7] We do this 1000 times and draw a histogram of interdecile differences

[7]Before pooling we normalize the two distributions to have the same mean by subtracting the ATE of 0.17 from the paired students' scores.

simulated over these 1000 replications. We then compare this histogram to the actual difference reported in Table 8. The p-value of the reported difference is the proportion of the histogram that lies to the right of the (positive) difference. For row 1, the difference is 2.67-2.61=0.06. Of the simulated differences under the null, 10% are larger than 0.06. The p-value of 0.06 is thus 10%. Similar calculations for row 2 and 3 yield p-values of 0.07 and 0.00, respectively. We therefore conclude that the reduction in dispersion induced by CAL is statistically significant.

We also calculate what further reduction in dispersion could be achieved with improved pairing. To this effect, we construct counterfactual distributions of math scores with negative assorting. This is achieved as follows. We first obtain predicted math scores for negatively assorted pairs following the methodology already described in the previous sub-section. By construction, the distribution of predicted scores has a smaller variance than actual scores because it omits the random variation contained in the residuals. In order to produce a counter-factual distribution that can be compared to the sample distributions presented in Table 8, we need to 'add' the error term back in. This is achieved by adding the residuals from regression (6) to the counter-factual predictions with improved pairing. We compare the resulting hypothetical distribution to the control population. Point estimates indicate that improved pairings generates a further – albeit small – reduction in the interdecile range of math scores. Applying the same permutation method as before to test whether the difference is significant, we find that it is not significant for all interdecile ranges reported in Table 8 – although it is borderline significant (p-value of 0.16) for the 90-10 interdecile range. These findings therefore do not suggest that negative assorting students would increase dispersion in math scores relative to random pairing – and may even reduce it.

6 Conclusion

We have conducted a large scale randomized controlled trial to investigate peer effects in learning. Identification of peer effects relies on three levels of randomization. We randomly assign schools to a treatment that successfully improves math learning. Within treated schools, students take the treatment either individually or in pairs. Finally, paired students are assigned a peer at random from the class population. In the methodological section, we show that this experimental

designs improves on earlier designs commonly used in the literature on peer effects in learning, such as paired designs used by Sacerdote (2001), Lyle (2007, 2009) and Shue (2012). We also avoid some of the pitfalls of paired designs discussed for instance in Guryan et al. (2009).

Our findings can be summarized as follows. Except for the first finding which confirms Mo et al. (2014), the others are all original to this paper.

1. In the Chinese rural schools we studied, computer assisted learning (CAL) leads to an average 0.17 standard deviation improvement in math scores among primary school students.

2. This average effect is the same whether students take CAL individually or in pairs.

3. There is no evidence of convergence in math scores among students who take CAL individually.

4. Among paired students, poor performers benefit more from CAL when they are paired with good performers.

5. Average performers benefit equally irrespective of who they are paired with.

6. Good performers benefit more from CAL when paired with poor performers.

Taken together, these findings allow us to conclude that (1) computer assisted learning improves math test scores in Chinese rural schools and that (2) paired treatment improves the beneficial effects of treatment for poor performers when they are paired with high performers, without hurting the performance of others. The second finding is similar to that reported by Booij, Leuven and Oosterbeek (2014) in the context of tutorial groups for university students.

One of the concerns at the onset of this experiment was that CAL could widen the knowledge gap between weak and strong students. This is not what we find. We test whether CAL treatment reduces the dispersion in math scores relative to controls, and we find statistically significant evidence that it does. We also demonstrate that the beneficial effects of CAL could potentially be strengthened, without significant increase in the dispersion of scores, if weak students are systematically paired with strong students during treatment. To our knowledge, this is the first time that a school intervention has been identified in which peer effects unambiguously

help poor student performers catch up with the rest of the class, without imposing any learning cost on other students. The treatment is good for both efficiency and equity.

We are not claiming that similar effects would be obtained by pairing students in other ways, for instance, as roommates. The treatment tested here may have stronger peer effects because it creates an environment that naturally induces students to interact. Roommates and other groups, on the other hand, may decide not to interact, as indicated for instance in the work of Carrel, Sacerdote and West (2013).

References

[1] Angrist, J. D., & Lang, K. (2004). "Does school integration generate peer effects? Evidence from Boston's Metco Program". *American Economic Review*, 1613–1634.

Bifulco, R., J. M. Fletcher, &S. L. Ross (2011). "The Effect of Classmate Characteristics on Post-Secondary Outcomes: Evidence from the Add Health". *American Economic Journal: Economic Policy*, 3(1), 25–53.

Bruhn, M., & D. McKenzie. (2009). "In pursuit of balance: Randomization in practice in development field experiments". *American Economic Journal: Applied Economics*, 1(4), 200–232.

Caeyers, B. (2013). Social Networks, Community-Based Development and Empirical Methodologies. Ph.D. thesis, University of Oxford Department of Economics.

Carrell, S. E., Fullerton, R. L., & West, J. E. (2009). "Does Your Cohort Matter? Measuring Peer Effects in College Achievement". *Journal of Labor Economics*, 27(3), 439–464.

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). "From natural variation to optimal policy? The importance of endogenous peer group formation". *Econometrica*, 81(3), 855–882.

CNBS [China National Bureau of Statistics]. (2011). *China National Statistical Yearbook, 2011*. China State Statistical Press: Beijing, China.

CNBS [China National Bureau of Statistics]. (2013). *China National Statistical Yearbook, 2013*. China State Statistical Press: Beijing, China.

Duflo, E., P. Dupas, & M. Kremer. (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". *American Economic Review*, 101(5), 1739–74.

Fafchamps, M., & P. Vicente. (2013). "Political Violence and Social Networks: Experimental Evidence from a Nigerian Election". *Journal of Development Economics*, 101, 27-48.

Fafchamps, M, A. Vaz, & P. Vicente. (2014). "Voting and Peer Effects: Evidence from a Randomized Controlled Trial". Stanford University (mimeograph).

Fletcher, J. M. (2010). "Social Interactions and Smoking: Evidence using Multiple Student Cohorts, Instrumental Variables, and School Fixed Effects". *Health Economics*, 19(4), 466–84.

Graham, B. S. (2008). "Identifying social interactions through conditional variance restrictions". *Econometrica*, 76(3), 643–660.

Guryan, J., K. Kroft, & M. Notowidigdo (2009). "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments". *American Economic Journal: Applied Economics*, 1(4), 34–68.

Hoxby, C. M., & G. Weingarth. (2005). Taking race out of the equation: School reassignment and the structure of peer effects. Working paper.

Kojima, F., & M. Utku Unver. (2013). "The 'Boston' School Choice Mechanism". *Economic Theory* (forthcoming)

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

Lai, F., L. Zhang, Q. Qu, X. Hu, Y. Shi, M. Boswell, & S. Rozelle. (2012). "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Public Schools in Rural Minority Areas in Qinghai, China." REAP working paper.

Lai, F., L. Zhang, Q. Qu, X. Hu, Y. Shi, M. Boswell, & S. Rozelle (2013). "Computer Assisted Learning as Extracurricular Tutor? Evidence from a Randomized Experiment in Rural Boarding Schools in Shaanxi". *Journal of Development Effectiveness*, 5(2), 208-231.

Lyle, D. (2007). "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point". *Review of Economics and Statistics*, 89(2), 289–299.

Lyle, D. (2009). "The Effects of Peer Group Heterogeneity on the Production of Human Capitalat West Point". *American Economic Journal: Applied Economics*, 69–84.

Manski, C.F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem". *Review of Economic Studies*, 60(3), 531-42.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., & Rozelle, S. (2014). "Integrating computer-assisted learning into a regular curriculum: evidence from a randomised experiment in rural schools in Shaanxi". *Journal of Development Effectiveness*, 6(3), 300–323.

Mo, D., L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell, & S. Rozelle (2013). "The Persistence of Gains in Learning from Computer Assisted Learning: Evidence from a Randomized Experiment in Rural Schools in Shaanxi Province". REAP working paper.

Booij, Adam S., Edwin Leuven and Hessel Oosterbeek (2014). "The Effect of Ability Grouping in University on Student Outcomes". University of Amsterdam

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics*, 116(2), 681–704.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? *Handbook of the Economics of Education*, 3, 249–277.

Shue, K. (2012). "Executive Networks and Firm Policies: Evidence from the Random Assignment of MBA Peers". Working Paper.

Vigdor, J., & Nechyba, T. (2007). Peer effects in North Carolina public schools. *Schools and the Equal Opportunity Problem*, MIT Press.

Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross-Section and Panel Data*. MIT Press.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics*, 85(1), 9–23.

Table 1. Balance between CAL school students and control school students, students who were paired and who sat alone in CAL classes, and between students who were assigned to a high achieving or a low achieving peer before attrition.

		Independent variables					
		CAL treatment (1=yes; 0=no)		Pair status (1=had a peer; 2=sat alone)		Standardized baseline math test score of the peer - class mean score (SD)	
		(1)	(2)	(3)	(4)	(5)	(6)
		Coef	S.E.	Coef	S.E.	Coef	Simulated P-values
[1]	Standardized own math test score - class mean score (SD)	0.00	0.00	0.04	0.07	-0.03	0.28
[2]	Boy (1=yes;0=no)	0.00	0.01	-0.01	0.03	0.00	0.43
[3]	Only Child (1=yes, 0=no)	0.01	0.03	0.03	0.04	0.00	0.45
[4]	Had computer experience before the program (1=yes;0=no)	0.00	0.03	0.07	0.05	0.00	0.48
[5]	Mother is illiterate (1=yes; 0=no)	0.00	0.01	0.02	0.02	0.01	0.21
[6]	Father is illiterate (1=yes; 0=no)	0.01	0.00	0.00	0.02	0.00	0.36

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

The test aims to present information about balance across the three different types of treatments in our experiment. The tests regress the variables listed on the left (each at a time) on the dummy variable of treatment status, the dummy variables of the pairing or the baseline math performance of the peer.

Table 2. Comparisons of attrition between the CAL school students and control school students, students who were paired and who sat alone in CAL classes, and between students who were assigned to a high achieving or a low achieving peer

Dependent variable: attrition (1=students attrited; 0=students remained in the sample)				
		(1)	(2)	(3)
[1]	CAL treatment (1=yes; 0=no)	-0.00		
		(0.01)		
[2]	Pairing status (1=had a peer;		-0.00	
	0=alone)		(0.02)	
[3]	Standardized baseline math score of			-0.00
	the peer - class mean score (SD)			(0.01)
[4]	Observations	7,881	3,852	3,675
[5]	R-squared	0.000	0.000	0.000

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

The test aims to show whether attrition rates are different among the groups defined by the three different types of treatment. The test regresses attrition status on the different treatment variable.

Table 3. Balance between CAL school students and control school students, students who were paired and who sat alone in CAL classes, and between students who were assigned to a high achieving or a low achieving peer after attrition

		Independent variables					
		CAL treatment (1=yes; 0=no)		Pair status (1=had a peer; 2=sat alone)		Standardized baseline math test score of the peer - class mean score (SD)	
		(1)	(2)	(3)	(4)	(5)	(6)
		Coef	S.E.	Coef	S.E.	Coef	Simulated P-values
[1]	Standardized own math test score - class mean score (SD)	0.00	0.00	0.04	0.07	-0.03	0.24
[2]	Boy (1=yes;0=no)	0.00	0.01	-0.02	0.03	0.01	0.21
[3]	Only Child (1=yes, 0=no)	0.01	0.03	0.02	0.04	0.00	0.46
[4]	Had computer experience before the program (1=yes;0=no)	0.00	0.03	0.07	0.06	0.00	0.45
[5]	Mother is illiterate (1=yes; 0=no)	0.00	0.01	0.02	0.02	0.01	0.11
[6]	Father is illiterate (1=yes; 0=no)	0.01	0.01	0.00	0.02	0.00	0.31

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

The test aims to present information about balance across the three different types of treatments in our experiment. The tests regress the variables listed on the left (each at a time) on the dummy variable of treatment status, the dummy variables of the pairing or the baseline math performance of the peer.

Table 4. The impact of the CAL treatment, the pairing status and the types of peer on own evaluation math score

	Dependent variable: Own standardized evaluation math score (SD)	[1]	[2]	[3]	[4]
[1]	Own standardized baseline math score (SD)			0.47***	0.50***
				(0.02)	(0.02)
[2]	Class mean of the standardized baseline math score (SD)	0.62***	0.63***	0.18***	0.17***
		(0.06)	(0.06)	(0.04)	(0.04)
[3]	CAL treatment (1=yes; 0=no)			0.17*	0.17*
				(0.09)	(0.09)
[4]	CAL treatment * (own score - class mean)[a]			0.00	0.00
				(0.08)	(0.09)
[5]	Being paired in CAL classes (1=yes; 0=no)			0.03	0.02
				(0.09)	(0.09)
[6]	Being paired * (own score - class mean)	0.47***	0.49***	0.02	0.04
		(0.02)	(0.02)	(0.09)	(0.09)
[7]	[Being paired * (own score - class mean)]^2		0.03**		0.04***
			(0.01)		(0.01)
[8]	Being paired * (peer score - class mean)[b]	0.02	0.01	0.02	0.01
		(0.01)	(0.02)	(0.02)	(0.02)
[9]	[Being paired * (peer score - class mean)]^2		-0.01		-0.01
			(0.01)		(0.01)
[10]	Being paired * (own score - class mean) * (peer score - class mean)	-0.04**	-0.03*	-0.04**	-0.03*
		(0.02)	(0.02)	(0.02)	(0.02)
[11]	Constant	0.20***	0.19***	0.00	-0.01
		(0.03)	(0.03)	(0.03)	(0.03)
[12]	Observations	3,524	3,524	7,536	7,536
[13]	R-squared	0.28	0.283	0.287	0.291

* significant at 10%; ** significant at 5%; *** significant at 1%. Robust standard errors in parentheses clustered at class level.

The tests aim to show how the CAL treatment, the pairing status and the types of peer affect own evaluation math score. The tests regress own evaluation math score on the variables listed on the left.

[a] The variable of "own score" refers to own standardized baseline math score (SD) and the variable of "class mean" refers to class mean of the standardized baseline math score (SD).

[b] The variable of "peer score" refers to the standardized baseline math score of the peer (SD).

Table 5. Predicted own evaluation math scores of students with high or low achieving peers using the regression model excluding the quadratic terms of test scores

		Peer score - class mean= -2	Peer score - class mean= -1	Peer score - class mean= 0	Peer score - class mean= 1	Peer score - class mean= 2	P-value (difference between columns 1 and 5)
		[1]	[2]	[3]	[4]	[5]	[6]
[1]	Own score - class mean= -2	-0.95	-0.85	-0.77	-0.69	-0.63	0.02
[2]	Own score - class mean= -1	-0.49	-0.42	-0.37	-0.33	-0.31	0.02
[3]	Own score - class mean= 0	0.03	0.06	0.07	0.08	0.07	0.21
[4]	Own score - class mean= 1	0.59	0.59	0.57	0.55	0.51	0.35
[5]	Own score - class mean= 2	1.21	1.18	1.13	1.07	0.99	0.14

The variable of "own score" refers to own standardized baseline math score (SD) and the variable of "class mean" refers to class mean of the standardized baseline math score (SD). The variable of "peer score" refers to the standardized baseline math score of the peer (SD).

Table 6. Predicted evaluation math test scores of students with high or low achieving peers using regression model including the quadratic terms of test scores

		Peer score - class mean=-2	Peer score - class mean=-1	Peer score - class mean=0	Peer score - class mean=1	Peer score - class mean=2	P-value (difference between column 1 and 5)
		[1]	[2]	[3]	[4]	[5]	[6]
[1]	Own score - class mean= -2	-1.02	-0.92	-0.82	-0.72	-0.61	0.06
[2]	Own score - class mean= -1	-0.48	-0.42	-0.36	-0.3	-0.24	0.08
[3]	Own score - class mean= 0	0.05	0.07	0.09	0.11	0.13	0.46
[4]	Own score - class mean= 1	0.59	0.57	0.54	0.52	0.5	0.38
[5]	Own score - class mean= 2	1.13	1.06	1.00	0.93	0.87	0.18

The variable of "own score" refers to own standardized baseline math score (SD) and the variable of "class mean" refers to class mean of the standardized baseline math score (SD). The variable of "peer score" refers to the standardized baseline math score of the peer (SD).

Table 7. Difference in predicted evaluation math test scores between control students and students that were paired

		Predicted evaluation math score of the control school students (without CAL)	Difference between the control school students and the paired students in CAL schools	Peer score - class mean= -2	Peer score - class mean= -1	Peer score - class mean= 0	Peer score - class mean= 1	Peer score - class mean= 2
		[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1]	Own score - class mean= -2	-0.88	Difference in scores (SD)	-0.07	0.03	0.11	0.19	0.25
			P-value	0.35	0.84	0.38	0.08	0.03
[2]	Own score - class mean= -1	-0.43	Difference in scores (SD)	-0.06	0.01	0.06	0.10	0.12
			P-value	0.64	0.72	0.16	0.03	0.01
[3]	Own score - class mean= 0	0.02	Difference in scores (SD)	0.01	0.04	0.05	0.06	0.05
			P-value	0.34	0.13	0.05	0.02	0.02
[4]	Own score - class mean= 1	0.48	Difference in scores (SD)	0.11	0.11	0.09	0.07	0.03
			P-value	0.08	0.08	0.14	0.32	0.59
[5]	Own score - class mean= 2	0.93	Difference in scores (SD)	0.28	0.25	0.20	0.14	0.06
			P-value	0.08	0.13	0.33	0.81	0.76

The variable of "own score" refers to own standardized baseline math score (SD) and the variable of "class mean" refers to class mean of the standardized baseline math score (SD). The variable of "peer score" refers to the standardized baseline math score of the peer (SD).

Table 8.Interdecile ranges of own evaluation math scores among the control school students and the paired students in CAL schools

		Control students	Paired students	Paired students in optimal matching	P-value for difference [1] - [2]	P-value for difference [2]- [3]
	Interdecile ranges	[1]	[2]	[3]	[4]	[5]
[1]	Q9 - Q1	2.67	2.61	2.57	0.10	0.16
[2]	Q8 - Q2	1.78	1.73	1.71	0.07	0.40
[3]	Q7 - Q3	1.19	1.08	1.08	0.00	0.54

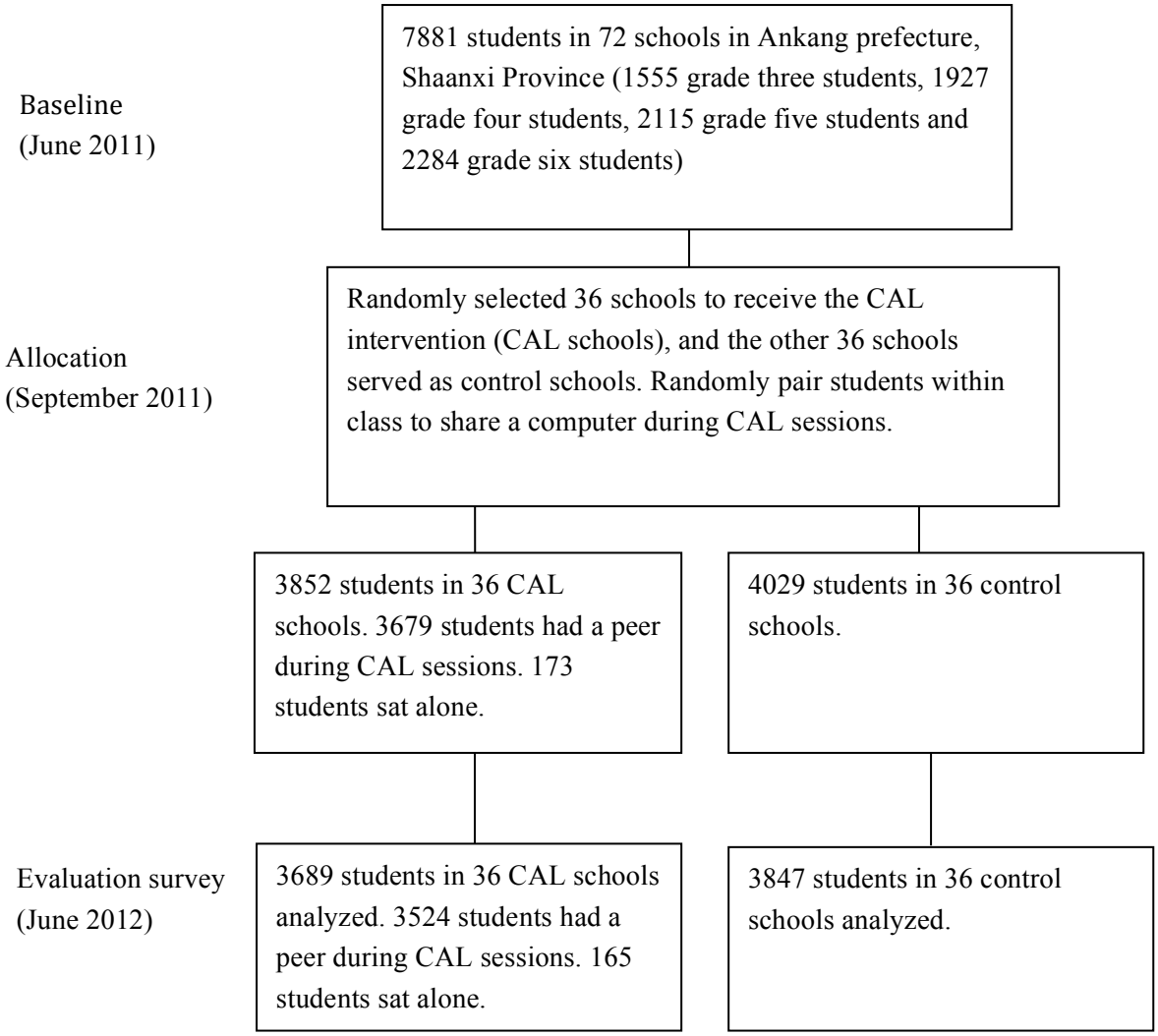

Figure 1: Experiment Profile

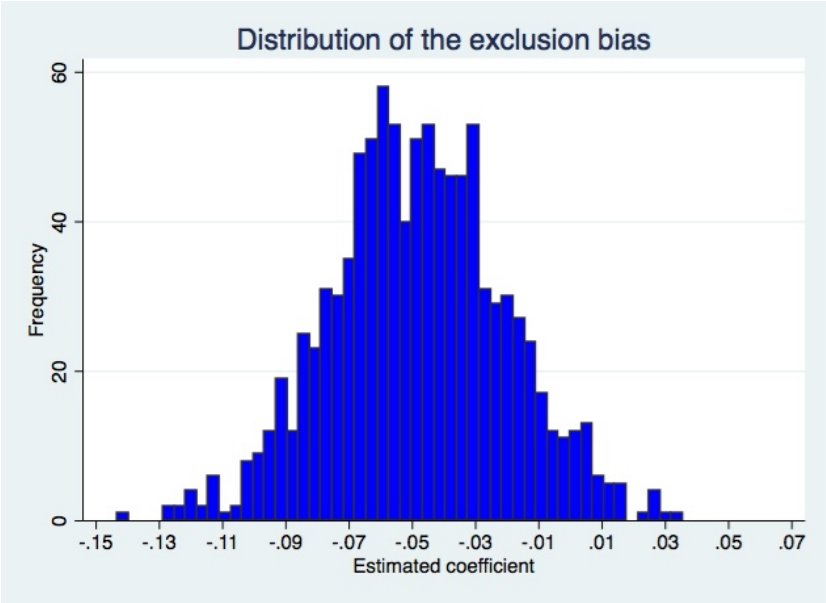

Figure 2. Histogram of $\hat{\hat{\beta}}_2$ under the null for baseline math scores

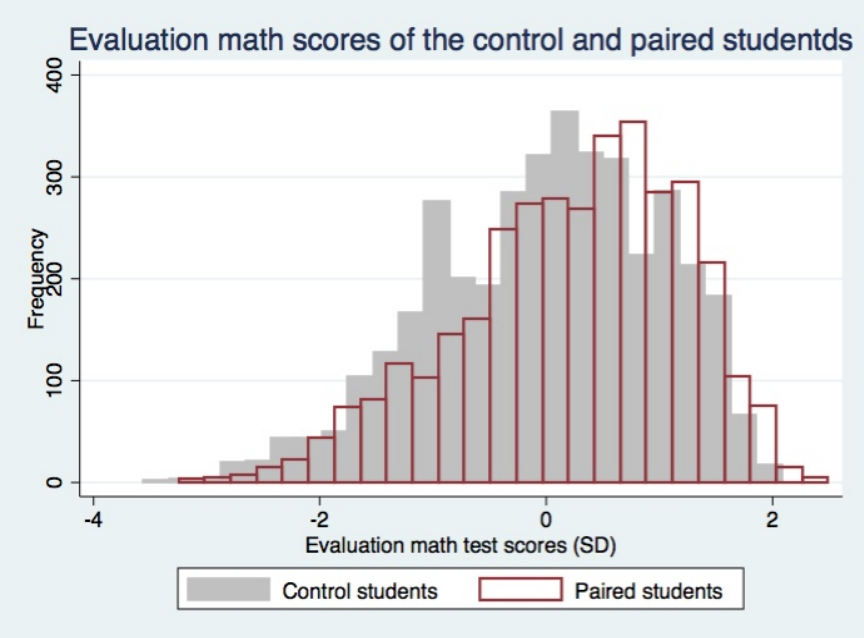

Figure 3. Comparison of own standardized evaluation math score between the control schools students and the paired students in CAL treatment schools

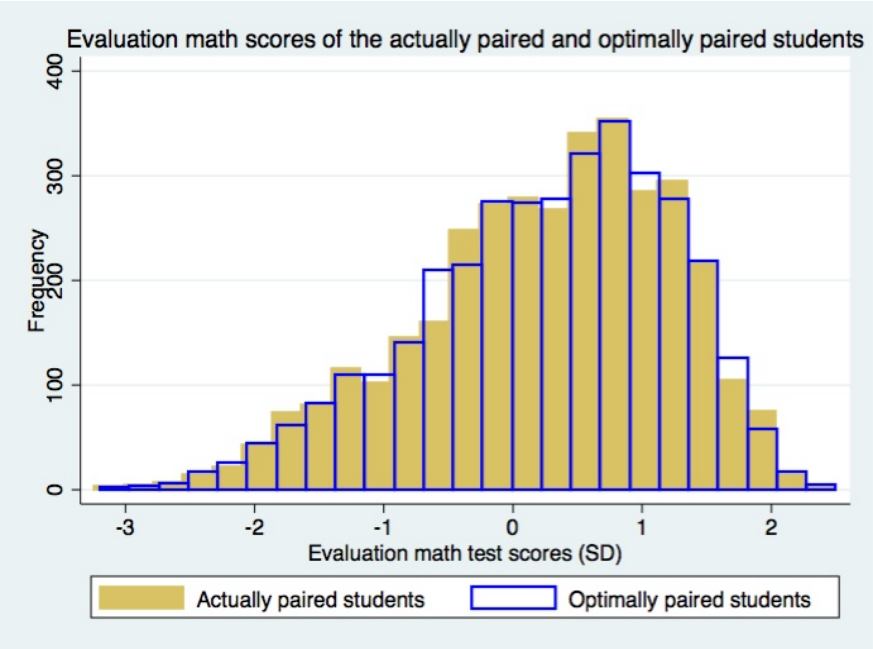

Figure 4. Comparison between the own standardized evaluation math score of the actually paired students in CAL treatment schools and the predicted own standardized evaluation math score of the optimally paired students

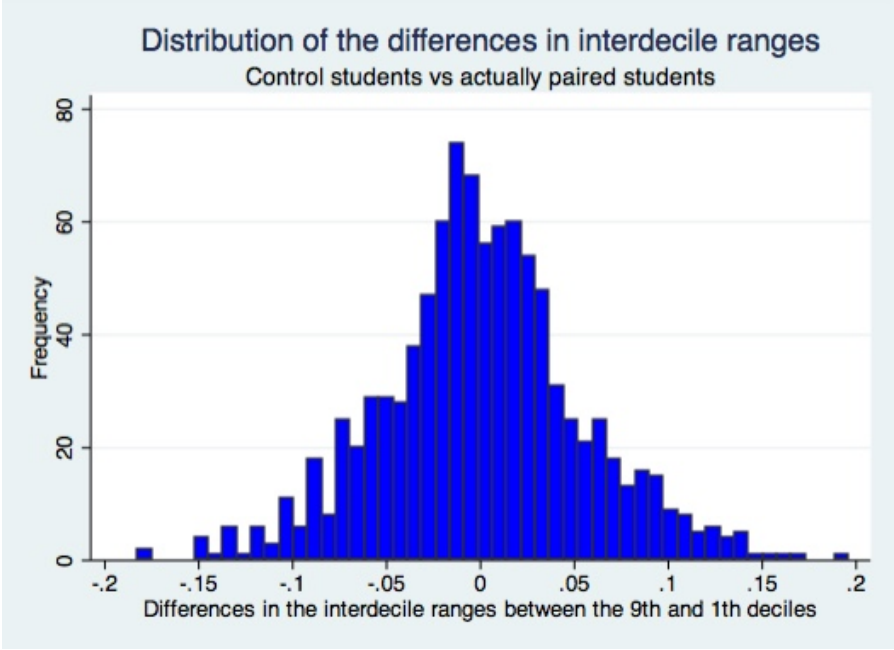

Figure 5. The simulated distribution of the difference in the interdecile ranges between the control school students and the paired students in the CAL treatment schools under the null that the score dispersion of the two groups is the same