Journal of Public Economics 111 (2014) 29–45

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Encouraging classroom peer interactions: Evidence from Chinese migrant schools☆

Tao Li [a,*], Li Han [b], Linxiu Zhang [c], Scott Rozelle [d]

[a] University of Macau, Faculty of Social Sciences, Av. Padre Tomas Pereira, Taipa, Macau
[b] Hong Kong University of Science and Technology, Hong Kong
[c] Chinese Academy of Sciences, China
[d] Stanford University, United States

ARTICLE INFO

Article history:
Received 27 December 2012
Received in revised form 20 December 2013
Accepted 27 December 2013
Available online 9 January 2014

ABSTRACT

In a randomized trial conducted with primary school students in China, we find that pairing high and low achieving classmates as benchmates and offering them group incentives for learning improved low achiever test scores by approximately 0.265 standard deviations without harming the high achievers. Offering only low achievers incentives for learning in a separate trial had no effect. Pure peer effects at the benchmate level are not sufficiently powerful to explain the differences between these two results. We interpret our evidence as suggesting that group incentives can increase the effectiveness of peer effects.

1. Introduction

The educational outcomes of low achieving students may improve if they are moved from a low achieving peer group to a higher achieving one through school integration policies such as detracking, school busing or housing vouchers (see the reviews of the peer effects literature by Epple and Romano, 2011; Sacerdote, 2011). Alternatively, their educational outcomes may improve if they experience more positive interactions with higher achieving peers in their *current* peer group — a hypothesis that has so far received little attention.

There are several reasons why we want to study this hypothesis. First, stimulating positive interactions between students from different backgrounds is arguably one of the ultimate goals of school integration policies. Mixing together a diversity of students within schools and classrooms is only one means for facilitating this purpose. Furthermore, altering peer groups in this way is both expensive and time-consuming, and yet there is no guarantee that it will automatically produce the desired peer interactions for low achieving students, who usually come from disadvantaged social backgrounds. Carrell et al. (2011) demonstrated a case in which randomly assigned higher achieving peers failed to benefit low achievers, and the most plausible explanation appeared to be the lack of interactions between these two groups. The lack of interactions between white and black students in officially integrated schools is also well documented (e.g., Echenique and Fryer, 2007). Finally, even if school integration policies successfully induce high and low achieving students to interact with each other, enhancing the quality of these interactions is still beneficial.

The way classrooms are typically managed in China offers us a convenient opportunity to study the above hypothesis. Traditional pair and row seating is the predominant classroom layout in China. Benchmate pairs typically sit next to each other throughout a semester and frequently interact with each other on a daily basis. By strategically reshuffling benchmates, student-level peer interactions could be influenced without the need to alter school or classroom composition.

Our *peer incentive experiment*, the focus of this paper, was designed to estimate the effects of enhancing benchmate-level peer interactions between high and low achieving classmates on the academic performance of low achievers. The experiment was implemented in 44 classes from 11 migrant primary schools in Beijing. Based on baseline test scores, we randomly assigned half of the bottom twenty students to the treatment group, and the other half to the control group. The treatment included an opportunity component as well as an incentive component. The opportunity component was that each treated student was randomly assigned to one of the top ten performing classmates as a benchmate for a semester. The incentive component was that the top three benchmate pairs (that is the three benchmate pairs in which the treated students made the largest test-score gains over a semester) in each class would get a monetary reward. The purpose of such a tournament-based group incentive was to encourage benchmates to interact with each other in a way that would contribute to the weaker partner's academic performance.

By comparing the treatment with the control students in the same classes (i.e., a within-class evaluation design), we found a robust effect

☆ We would like to thank Weili Ding, Esther Duflo, Gigi Foster, Caroline Hoxby, Michael Kremer, Victor Lavy, Steven Lehrer, Patrick McEwan, Albert Park, Sujata Visaria and Hongliang Zhang, two anonymous reviewers and the seminar participants at Harvard, Stanford, Peking University, NEUDC, Oxford and Chinese University of Hong Kong for their helpful comments on the drafts of this paper. Financial support from Ford Foundation and the University of Macau (Grant SRG009-FSH13-LT) is gratefully acknowledged.

* Corresponding author.
E-mail address: litao@post.harvard.edu (T. Li).

0047-2727/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jpubeco.2013.12.014

of approximately 0.265 standard deviations (s.d.) in the low achievers' evaluation test scores using various estimation strategies. This finding is significant because it clearly demonstrates that policymakers can make peer effects more effective than they would otherwise be without the time or expense associated with manipulating classroom or school composition.

Behind any economic intervention that employs a group incentive scheme there are many potential mechanisms that could be at play. In our peer incentive experiment, there are at least three potential mechanisms to consider: treated low achievers improved their scores (1) because of peer interactions stimulated by group incentives (2) because of their own desire to win rewards from their scores, or (3) simply because of having better opportunities to interact with a high-achieving classmate. Unpacking our main effects has both theoretical and practical implications.

For this purpose, we additionally ran a separate *individual incentive experiment* that studied the effects of offering low achievers exactly the same level of incentives for improving their scores (but did not match them up with a high-achieving peer) in 47 classes from 12 different migrant primary schools in Beijing. We also followed the peer effects literature (e.g., Sacerdote, 2001; Zimmerman, 2003) and estimated conventional, reduced-form *pure benchmate effects* (i.e., pure peer effects at the benchmate level) by exploiting two exogenous changes to benchmate composition in the peer incentive classes. Neither the individual incentive nor the pure benchmate effects are statistically distinguishable from zero. The evidence supports a straightforward interpretation of our primary finding: in our peer incentive experiment it is the group incentive, rather than either of the two alternative mechanisms, that made the peer effects more effective than they would have been otherwise.

At the class level, we randomly assigned 35 extra classes to be control classes. By comparing students from the experimental classes to their counterparts in the control classes with similar baseline test scores (i.e., an across-class evaluation design), we found a small and statistically insignificant spillover effect for the untreated students in the experimental classes, including the high achievers in the peer incentive classes. Our results suggest that encouraging peer interactions inside a given peer group may be a less controversial way to make use of peer effects because it brings about efficiency gains.

We nevertheless acknowledge that this paper has several important limitations. Because of the small number of schools involved in our study, we cannot discuss the effects that result from an entire school being treated. One could imagine that the culture of the school could change in a general way. A larger study involving school-level treatment would improve the external validity of our research. Another line of future research would be evaluating the long-term effects of our peer incentive treatment. The effects reported here were short term (one semester only); the long-term effects, if any, are unclear.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 presents a conceptual framework. Section 4 describes our programs and data. Section 5 describes the evaluation design and reports results from the peer and individual incentive experiments. Section 6 reports the estimation strategies and the results of the pure benchmate effects. Section 7 concludes the paper. Details of program implementation and some extra robustness checks are in the Appendix.

2. Related literatures

To the best of our knowledge, benchmate pairs are the smallest set of peer groups that the classroom peer effects literature has ever studied. Benchmate interactions are entirely voluntary. There are no assigned tutoring sessions. Teachers are not involved in the daily interaction process. These two characteristics make benchmate interactions fundamentally different from cooperative learning intensively studied by educational psychologists (Johnson and Johnson, 1997) or group studying and peer tutoring studied by other economists (Angrist et al., 2009; Blimpo, 2010).

Nearly all the previous empirical literature on educational peer effects focuses almost exclusively on the task of establishing whether peer effects exist by exploiting exogenous changes in peer group composition.[1] The reduced-form peer effects estimated in this way cannot be used to distinguish among externalities from different channels (Manski, 1993). With objectives similar to the work in this paper, several authors have recently tried to estimate peer effects emanating from different student behavior, such as student efforts (Cooley, 2009), the choice of college major (Giorgi et al., 2009) and classroom infractions (Kinsler, 2010). None of these papers, however, have explicitly studied peer interactions. To our knowledge only a few papers have attempted to do so. Relying on surveys and administrative data, Stinebrickner and Stinebrickner (2006, 2008) found that college roommate peer effects are most likely to arise through roommates influencing each other's time-use rather than through their interacting on academic matters. The paper by Carrell et al. (2011) is in spirit closer to ours. They found that high and low achieving peers may be reluctant to interact in schools, which might contribute to the poor academic performance of low achievers. However, unlike our experiment, their study was not designed to provide causal evidence of the effect of peer interactions on educational outcomes.

In another literature it has been well established that cash incentives are effective in stimulating peer interactions in workplaces (Hamilton et al., 2003; Boning et al., 2007; Chan et al., 2010). As far as we know, however, there is no parallel study on peer interactions in schools, except for two papers by Babcock et al. (2010), Babcock and Hartman (2011) that we will discuss below. This lack is a bit surprising because peer effects are considered to be a central input into the education production process (Epple and Romano, 2011). The absence of the use of cash incentives to encourage peer interactions cannot be explained by a lack of interest in using cash incentives in education. The use of cash incentives to solicit other types of socially desirable behavior in education has flourished in recent years, such as conditional cash transfer programs surveyed by Rawlings and Rubio (2005), teacher merit pay programs surveyed by Podgursky and Springer (2007), and randomized trials encouraging college student workout behaviors by Babcock et al. (2010) and Babcock and Hartman (2011). The latter two papers are similar in spirit to ours in that they examined the effects of cash incentives on randomly-assigned or self-selected student peer groups. Importantly, the two Babcock studies did not examine academic outcomes.

The segment of the literature that examines the use of cash incentives in school that is probably most relevant to our study is the set of studies that examine pay-for-grades programs (which we call individual incentive experiments in this paper). Over the past decade, a large number of such programs have been implemented around the world. Despite much enthusiasm, the estimated program effects on actual learning are still mixed (for a review, see Slavin, 2010). Evidence of the effect of pay-for-grades programs on secondary school students, usually those preparing for important high school exit exams, tend to be positive and significant (Mauldon et al., 2000; Spencer et al., 2005; Angrist and Lavy, 2009; Jackson, 2010). In contrast, evidence for primary school students is noisier (Kremer et al., 2009; Bettinger, 2012; Fryer, 2011). After analyzing the effects from the largest pay-for-grades experiment conducted in 261 American public schools, Fryer (2011) suggested that individual incentives tied to student test scores were not effective because students did not know how to improve learning on their own. Fryer's conclusion underscores the need to compare pay-for-grades programs with programs that not only pay for grades but

[1] An incomplete list of recent contributions include the following: Foster (2006), Ding and Lehrer (2007), Figlio (2007), Lyle (2007), Carrell et al. (2009), Carrell and Hoekstra (2010), Burke and Sass (2011), Gibbons and Telhaj (2011), Lavy and Schlosser (2011), Imberman et al. (2012), Lavy et al. (2012), etc.

also try to provide students with other educational resources (e.g., peer-level interactions — which are precisely what we do in this study).

3. A conceptual framework

The following framework is based on Cooley (2009), who explicitly studied peer effects emanating from student efforts.[2] Consider a low achieving student i and the student's high achieving peer j. Keeping inputs from parents and schools fixed, we can write the following test score production function for student i:

$$y_i = f\left(a_i, e_i; a_j, e_j\right) \tag{1}$$

where test scores y_i are affected by the student's characteristics a_i (including ability, race, gender, and family background), efforts e_i, peer characteristics a_j, and the effort of the student's peer e_j. The production function for j can be specified similarly.

We depart from the literature by assuming that e_i (and similarly e_j) includes two types of efforts (or two tasks), namely studying effort e_i^s (task 1) and cooperative (or interactive) effort e_i^c (task 2, the focus of this paper). These terms were borrowed from Cooley (2009). We distinguish between these two types of effort to emphasize their conceptual differences. Either of them can be underprovided, but for very different reasons.[3] Using this conceptualization, we can then provide an analytical framework for each of the three effects studied by this paper.

The first effect was the individual incentive treatment effect (i.e., the conventional pay-for-grades program effect). Student i can be paid with short-term cash incentives to improve test scores if she exerts higher e_i^s (task 1). The student does not exert high e_i^s without the short-term incentive because she otherwise lacks motivation, does not know the long-run returns to education, or severely discounts the future benefits of education (Angrist and Lavy, 2009; Bettinger, 2012; Fryer, 2011).

The second effect was a reseating effect or a pure peer effect. Following Manski (1993) and Cooley (2009), we distinguished between two types of peer effects. Student i can benefit from having a benchmate with "better" a_j and higher e_j^s. a_j has a direct impact on i's scores (contextual effect), while e_j^s increases y_i by stimulating higher e_i^s (endogenous effect).

The third effect was the peer incentive treatment effect. This treatment was a combination of reseating and an unconventional Leontief peer incentive contract, which based payment to each pair of benchmates on the low-achieving student's score y_i.[4] The payment was then shared by i and j equally. This group incentive contract can help stimulate e_i^c and e_j^c (task 2) by internalizing externalities, but at the same time it also directly stimulated e_i^s (task 1). Therefore, this treatment was a combination of three components: reseating, incentives for task 1 (for student i only), and incentives for task 2 (for both students). Comparing the effects of our peer incentive treatment to pure reseating effects and individual incentive effects can thus help distinguish the true effects of encouraging peer interactions.

We chose to use the Leontief group incentive contract instead of the conventional group incentive contract based on average group output $(y_i + y_j)/2$ for our peer incentive experiment.[5] The latter contract is a function of both y_i and y_j, so it directly stimulates e_j^s as well as e_i^s, which creates at least two additional complications for understanding the direct incentive effects on task 2, even if we only look at the effect on y_i. The first complication is a spillover (endogenous) effect. When y_j increases as a result of higher e_j^s, then y_j could have a spillover effect on y_i, even if both e_i^c and e_j^c are zero. For example, in Kremer et al. (2009), merit-based scholarships for girls had a positive effect on both boys and low scoring girls who were ineligible or unlikely to win scholarships. The second complication is the usual substitution effect in a multi-task setting. Student j may find it worthwhile to spend time on her own study (e_j^s) instead of interacting with student i (e_j^c). The overall impact of these two opposing forces on y_i is thus ambiguous, such that it is difficult to determine the extent to which peer interactions (in terms of e_i^c and e_j^c) are directly stimulated by the conventional group incentive contract. Another reason that we favor the Leontief incentive scheme is its policy relevance in targeting academically weak students.

4. The randomized trial

4.1. Background

There are approximately 150 million rural migrant workers in China. Because the government tightly controls family-based permanent migration, the children of migrant workers face systematic discrimination in China's cities. For example, migrant children in Beijing typically cannot attend local public schools unless they pay a high fee and/or complete substantial paperwork. Even if migrant children attend low cost, fee-based migrant schools, as in our study, they are not allowed to take either the high school or college entrance exam in Beijing. Authorities provide almost no resources to migrant schools and at times have been known to arbitrarily close the schools without warning. Although Beijing is one of China's richest cities, the migrant schools there more closely resemble schools in underdeveloped rural areas.

4.2. Treatment design

We refer to all the treated students as participating in either the *individual incentive experiment* or the *peer incentive experiment*. Each class either hosted exactly one of these two experiments (referred to as an experiment class) or did not host an experiment at all (referred to as a control class). Each experiment class had approximately ten treated students.

We will first introduce the individual incentive experiment. Treated students in the individual incentive experiment classes were offered a pay-for-grades incentive contract. We promised to make a payment of 100 RMB (approximately 13 U.S. dollars or between one-third and one-quarter of a semester's tuition) to the student with the greatest increase in test scores between the baseline test (taken in September 2009) and the evaluation test (taken in January 2010). The second and third place runner-ups were promised 50 RMB each. In total, we offered 200 RMB to be split among three winners from amongst the ten treatment students in each individual incentive class. We also promised a public ceremony and official certificates for the winners.

Our pay-for-grades incentive contract was a tournament. Existing pay-for-grades programs (reviewed in Slavin, 2010) typically reward all participating students based on absolute test scores or on whether students reach a certain target (i.e., a linear or piece-rate incentive contract). Although we used standardized tests for both the baseline and the evaluation tests, it is technically difficult to design tests in such a way that the difference between the two test scores can correctly measure the improvements resulting from our semester-long intervention, not to mention the difficulty of maintaining a uniform standard for students from different grades who would have to take different tests. The use of a tournament contract simplified the test design process. Moreover, because we implemented the tests and performed all the grading ourselves, an incentive contract based on absolute test scores would

[2] Cooley's framework is closely related to Manski (1993) and Brock and Durlauf (2001).

[3] Blimpo (2010) also distinguished between these two efforts. However, he did not consider pure peer effects operating through a_j and e_j^s on student i's test scores. Not considering pure peer effects distinguished his paper from the peer effects literature in the tradition of Manski (1993). The advantage of Blimpo (2010) was the ability to obtain closed-form equilibrium solutions under various pay-for-grades schemes.

[4] The incentives for both students effectively depend on $min\{y_i, ky_j\}$. We assume that k is a sufficiently large positive number such that $ky_j > y_i$. Note that even if we assume $k = 1$, y_j is likely to be larger than y_i because of initial talent difference.

[5] Blimpo (2010) used this type of contract to encourage group studying.

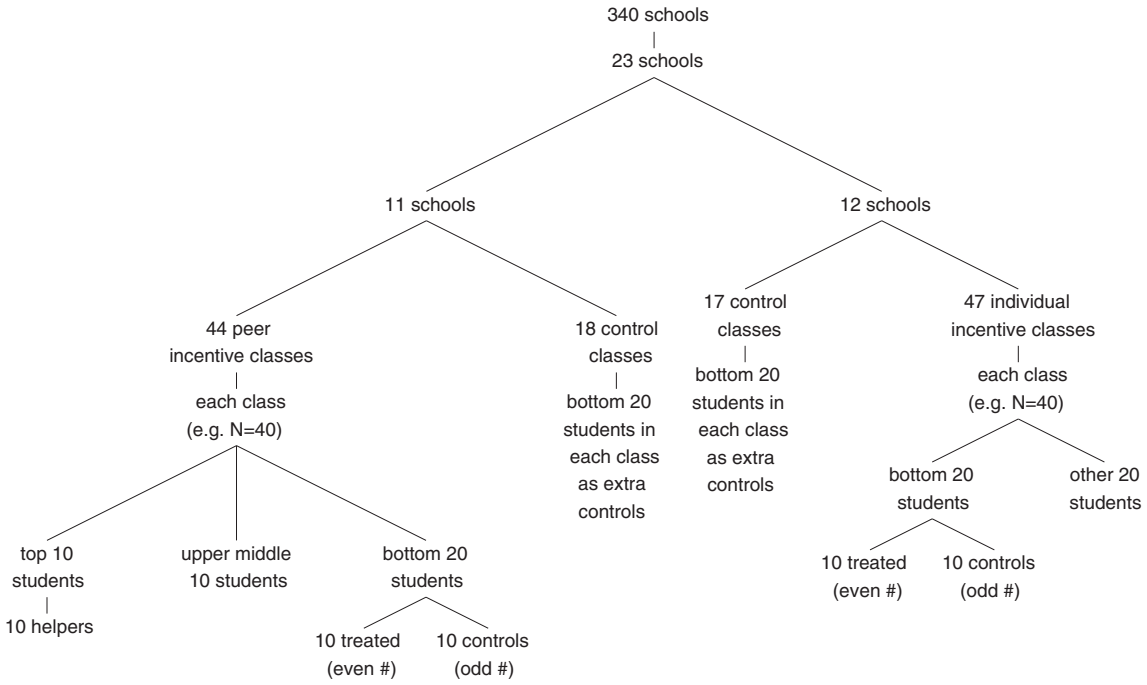


Fig. 1. Experiment design chart.

have made it difficult to convince the teachers and students ex ante that we would not increase the difficulty level of the tests or implement a stricter grading policy to minimize our payout. The use of a tournament contract also eliminated potential grading bias. Kremer et al. (2009), who designed and evaluated one of the few existing pay-for-grades programs in a developing country, also used tournament incentive contracts.

The ten treated students in each peer incentive class were offered the same tournament contract as the treated students in the individual incentive classes, such that the incentive effects are comparable. In addition, in the peer incentive class we assigned each of the top ten students in the class to serve as a benchmate for one of the treated students. To encourage peer interactions, we not only promised to award the three treated students with the greatest test score gains, but also promised to award their assigned (top student) benchmates with an equivalent cash prize. As a result, our budget for each peer incentive class was 400 RMB instead of 200 RMB. We gave no instructions regarding how benchmates should interact with each other.

4.3. Random assignment

The detailed school, class and individual assignment process is illustrated in Fig. 1. From a nearly complete list of 340 migrant schools in Beijing we randomly selected 23 schools to participate in our study. Our study focused on students in grades 3 through 6. Each grade typically had one to three classes. Within each school, we randomly selected four to six classes in our target grade range (grades 3 through 6). Each grade in each school had at least one class included in the study. We did not pick more than two classes from the same grade from the same school. We enrolled a total of 126 classes into our study. Every student in these classes participated in a baseline test (pre-test) and survey and an evaluation test (post-test).

Contingent on the baseline standardized test scores and information produced from the data collected during the survey, we randomly chose 12 schools to host the individual incentive experiment. The other 11 schools hosted the peer incentive experiment. After randomization, the 13 pre-treatment variables were balanced across these two groups of schools.[6]

In most schools, we randomly selected four classes to implement the assigned experiment; the remaining one or two classes served as control classes.[7] There were 44 peer incentive classes from 11 peer incentive schools, 47 individual incentive classes from the other 12 schools (individual incentive schools), and 35 control classes from all 23 schools. After randomization, the same 13 pre-treatment variables were well balanced among students in the three types of classes.[8]

In each experiment class, we initially ranked all students by their combined scores on the math and Chinese tests taken during the baseline. In a typical class of 40 students, we divided the class into quartiles: the top 10 students in quartile 1; the next 10 in quartile 2; and the poorest performing students (ranks 21 to 40) in the bottom two quartiles. In half of the experiment classes, we selected students in the bottom half with odd-numbered rankings (21, 23, …, 39) to receive treatment. The even-numbered students in the bottom half acted as control students. In the other half of the experiment classes, we selected the even-numbered students to act as treatment students instead and the odd-numbered students functioned as control students. The above randomization procedure ensured that the bottom treated group (BT) and bottom control group (BC) had almost identical size and baseline scores.

The above individual-level random assignment was the same for the peer and individual incentive classes with one difference. The top ten students in the peer incentive classes (T or Quartile 1) were randomly assigned as benchmates of the BT students, while the top ten students

[6] The pre-treatment variables included year of birth, whether a student had repeated grade, whether a student's father (or mother) was away from home for job reasons, the number of older brothers (or sisters), whether any of the older brothers (or sisters) attended high school, and two standard self-esteem questions. We followed the rerandomization methods described by Bruhn and Mckenzie (2009) — we continued randomizing until we obtained a sample with balanced characteristics. The detailed school-level assignment process and randomness checks are in Appendix A.

[7] In a few schools with only four classes, all classes implemented the assigned program.

[8] Detailed class-level randomness checks are omitted to save space (available upon request).

Table 1
Data summary.

	Class type		
	Peer inctv.	Individual inctv.	Control
Pre-test	0.004	−0.010	0.008
(Standardized test score)	(1.029)	(0.998)	(0.965)
Post-test	0.028	0.003	−0.0401
(Standardized test score)	(0.994)	(1.011)	(0.992)
Male (1/0)	0.565	0.554	0.549
(1 if male)	(0.496)	(0.497)	(0.498)
Treatment (1/0)	0.217	0.230	0
(1 for treatment)	(0.412)	(0.421)	(0)
N of schools	11	12	23
N of classes	44	47	35
N of students	1710	1789	1351
In grade 3	449	505	392
In grade 4	436	465	330
In grade 5	414	452	360
In grade 6	411	367	269
N of treated students	371 (226 boys)	411 (246 boys)	
In grade 3	91 (52 boys)	109 (72 boys)	
In grade 4	93 (57 boys)	104 (61 boys)	
In grade 5	96 (57 boys)	103 (61 boys)	
In grade 6	91 (60 boys)	95 (52 boys)	

Note: the sample corresponds to all the students who took our baseline survey and test. Standard deviations reported in parenthesis.

in the individual incentive classes did not participate actively in our experiment and there was no benchmate reshuffling in these classes.

4.4. Implementation

In late August 2009, using the methodology discussed above, we chose 23 schools and 126 classes from these schools to participate in our study. In early September, our enumerators administered a standardized multiple-choice test to these 126 classes at the beginning of the semester. This baseline test had two parts: math and Chinese. All classes in the same grade used the same test in all schools. Different grades used different tests. We additionally conducted a baseline survey that asked students to answer basic background questions about themselves and their families. All students in the same school received the test simultaneously. The exam was given in a 3-day interval across different schools, which are scattered throughout Beijing, a vast metropolitan area. Because students wrote down answers on the exam papers, essentially all exams were returned to us after the test. The chance of leaking is low. In each test room, one teacher and one enumerator acted as exam proctors, and one or two additional enumerators walked around the test rooms as monitors. The exams were graded by computer soon after the test was administered.

Our enumerators returned to the 23 schools in the middle of September and implemented the random assignment of control and treatment groups. The details of interacting with students/parents, including an extra round of parental communication with half of the treated students, are discussed in Appendix B. In early January 2010, our enumerators returned to the schools to conduct a standardized evaluation test. Shortly after the evaluation survey, awards and certificates were distributed to the students in an official ceremony held in each school. In total, 27,000 RMB (or approximately 4000 U.S. dollars) was distributed.

Table 2
Randomness checks of two evaluation designs.

	Peer incentive			Individual incentive		
	Control	Tr.	Diff.	Control	Tr.	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)
A: within-class design						
Pre-test	−0.747	−0.813	0.066	−0.820	−0.814	−0.006
			(0.071)			(0.059)
Male	0.614	0.609	0.005	0.579	0.599	−0.020
			(0.036)			(0.034)
N	376	371		418	411	
B: across-class design						
Pre-test	−0.786	−0.813	0.027	−0.786	−0.814	0.028
			(0.061)			(0.055)
Male	0.571	0.609	−0.039	0.571	0.599	−0.028
			(0.033)			(0.032)
N	587	371		587	411	

Note: This table checks the balance of two baseline variables in two evaluation designs using t-test. The samples used in the within-class design (panel A) are bottom students in experiment classes. The samples used in the across-class design (panel B) are bottom treated students in experiment classes and bottom half students in control classes. "Diff" refers to the difference between control and treatment groups. Standard errors reported in parenthesis. *p < 0.1, **p < 0.05, ***p < 0.01.

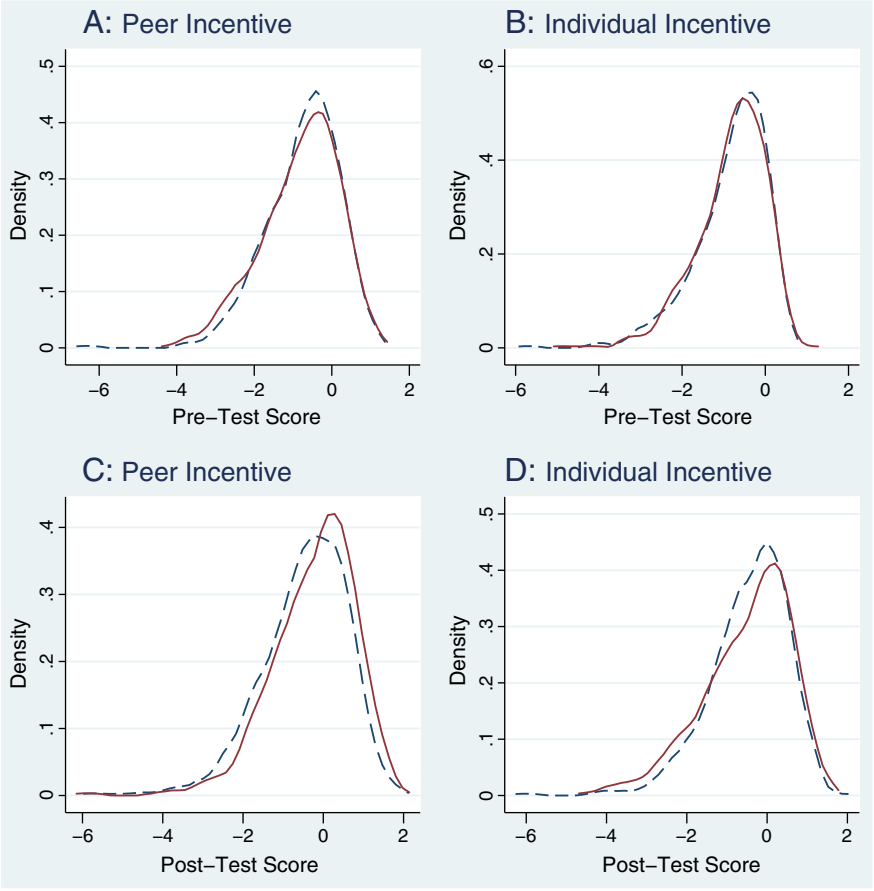
4.5. Data description

We focus on the students that took both tests (N = 4850).[9] Their data are described by different class types in Table 1. The variables *Pre-test* and *Post-test* represent the standardized scores from the baseline and evaluation tests, respectively.[10] The *Male* variable is a binary dummy variable for gender. There are slightly more males in the sample (approximately 55%), an imbalance frequently found in schools for students from rural areas in China. The *Grade* variable takes the value of 3, 4, 5, or 6. We choose not to include other background variables from the baseline survey because of the frequency of missing values. Our primary results remain robust if we include additional control variables with missing values.

Table 1 shows that there was little difference between the control and experiment classes in terms of the three pre-treatment variables. Table 1 additionally shows that treatment status was distributed evenly across the two types of experiment classes. Specifically, 21.7% of the students in peer incentive classes and 23.0% of the students in individual incentive classes were treated.

[9] The attrition rate (approximately 10%) is quite uniform across the control and different treatment groups (11% vs. 10% in peer incentive within-class experiment, with a p-value of 0.68; 9% vs. 8% in individual incentive within-class experiment, with a p-value of 0.59). Our interviews with the teachers corroborated this pattern, which is most likely due to student absence and the high mobility of migrant children.

[10] Standardization (for math and Chinese separately) was done in each grade/year combination because different grades used different tests in different years. Then, standardized math and Chinese scores (z-scores) are averaged to produce the pre and post test scores.



Note: Our sample is the bottom students in the within-class design (bottom treated students or BT, and their within-class controls BC). The solid lines refer to the BT groups. The dashed lines refer to the BC groups.

Fig. 2. Density distributions of test scores in the within-class design. Note: Our sample is the bottom students in the within-class design (bottom treated students or BT, and their within-class controls BC). The solid lines refer to the BT groups. The dashed lines refer to the BC groups.

5. Peer versus individual incentive effects

In this section, we describe the evaluation design and report on the estimation results for the peer incentive effects and individual incentive effects. Pure benchmate effects are reported in the next section.

5.1. Evaluation sample and randomness checks

Almost all previous studies randomized at the school level, including one of the largest pay-for-grades experiments to date (reported in Fryer, 2011). That study included 261 American public schools across four school districts for a total of 123 treatment schools and 138 control schools. Although the total number of students participating in the study was large, approximately 38,000, the number of effective experimental units was only 261. The Girl Scholarship Experiment in Kenya (Kremer et al., 2009) had a sample of 127 schools from two districts, with attrition complicating estimation in one district. Blimpo's (2010) pay-for-grades experiments in Benin had a sample of 100 schools. However, because he divided his schools into four approximately equal groups (three separate treatment groups and one control group), the effective sample size for each of his experiments was 44, 44, and 56, with statistical power estimated to be approximately 0.85 (Blimpo, 2010, Table 3).

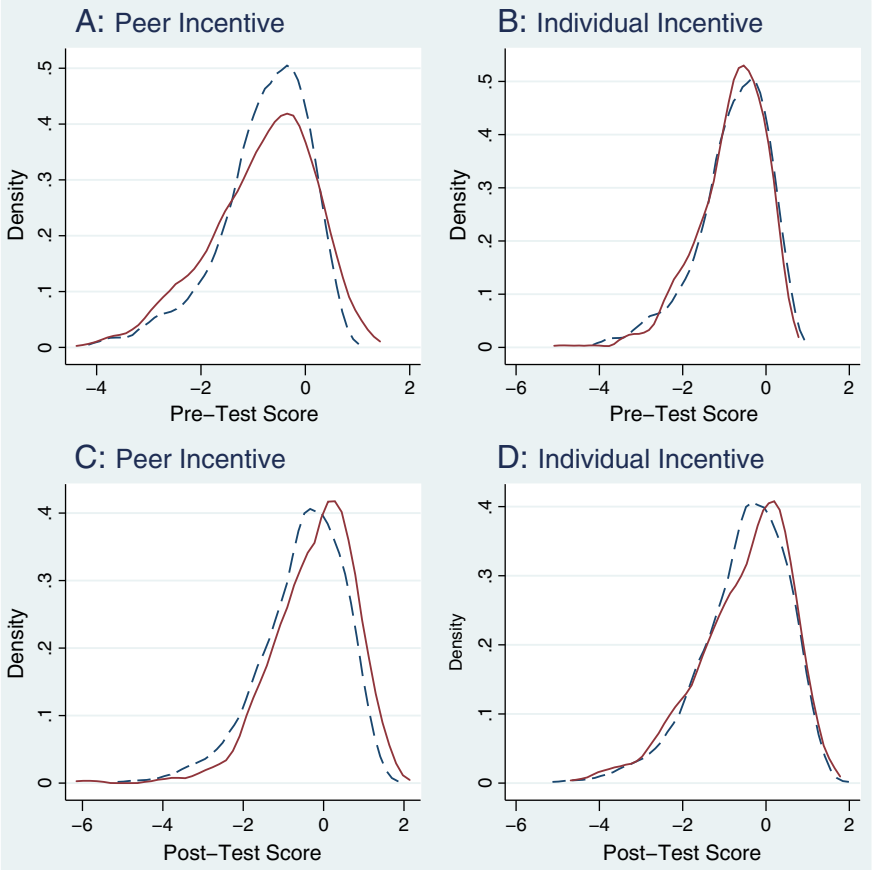
One notable feature of our intervention is that it has a relatively large sample size because of the use of individual-level or class-level randomization. This can be illustrated using the peer incentive experiment. In our within-class design, the unit of study is the individual student. The sample size is 747, with 371 treatment students and 376 control students (Table 2 Panel A). In our across-class design, the unit of study is the class. The sample size is 79, with 44 treatment classes and 35 control classes.

The power for our within-class design is at least 0.95 for a treatment effect of 0.25 standard deviations (same below).[11] Similarly, the power for our across-class design is at least 0.9. Also note that unlike Kremer et al. (2009) and Blimpo (2010), our randomization explicitly balanced important baseline variables. In our within-class design, if we assume that the pre-test scores can explain 50% of the variation in the outcome, the calculated power is approximately 0.99 if this variable is controlled for when we estimate the treatment effects.

Because of the use of within-class and cross-class/within-school designs, we are confident that the number of schools that we use in this study is adequate for us to estimate the effect of the treatment.[12] Our

[11] Suppose that blocking on classroom explains 30% of the variation in the outcome. Assume the effect size variability to be 0.01, the treatment effect to be 0.25 standard deviations, and the number of students to be assigned per class to be 20. The power calculations of our multi-site (blocked) person randomized trials are performed using *Optimal Design* software. Details are available upon request.

[12] Other researchers cannot take advantage of within-school randomization largely because of the political sensitivity of rewarding some students in a grade for their achievement and not others (Fryer, 2011), a concern that is less important in our migrant school context.



Note: The sample is the bottom treated students (BT) in experiment classes and bottom half students in control classes. The solid lines refer to the BT groups. The dashed lines refer to the bottom half students in control classes.

Fig. 3. Density distributions of test scores in the across-class design. Note: The sample is the bottom treated students (BT) in experiment classes and bottom half students in control classes. The solid lines refer to the BT groups. The dashed lines refer to the bottom half students in control classes.

compact design, however, comes with a cost. With a small set of schools, it is possible that external validity is more of an issue.

Fig. 2 (and Fig. 3) panels A and B show that the baseline test scores are balanced between the control and treatment groups in our within-class design and across-class design, respectively.

Table 2 shows the descriptive statistics and randomization checks for both the within-class design (panel A) and the across-class design (panel B) of the individual and peer incentive experiments. In both designs, the differences between the pre-treatment variables in the control and treatment groups were quite small and were not significantly different from each other at the 10% level.

5.2. Evaluation design

The main empirical strategy is captured by the following OLS regression, which can be applied to both the within-class and across-class designs:

$$\text{post-test}_i = \alpha + \beta Treatment_i + \mathbf{X_i}\gamma + \epsilon_\mathbf{i} \quad (2)$$

where *Post-test* and *Treatment* are the dependent and key explanatory variables described above and in Table 1. The matrix $\mathbf{X_i}$ includes our two control variables, *Pre-test* and *Male*, as well as a set of class dummies. Standard errors throughout the paper are clustered at the class level unless stated otherwise. To avoid multicollinearity, we use grade instead of class dummies in the across-class design. β captures the treatment effect.

Because each treated student had a within-class control with a nearly identical pre-test score, we can estimate the Average Treatment Effect (ATE) using one-to-one matching based on pre-test ranking in the same class (i.e., we can match the 40th treatment student with the 39th control student, and then the 38th treatment student with the 37th control student and so on).

5.3. Primary results

Panels C and D in Fig. 2 provide graphical evidence of the treatment effects using the within-class design. The graph indicates a positive effect (approximately a quarter of a standard deviation) from the peer incentive treatment on standardized test scores but no effect from the individual incentive treatment. Panels C and D in Fig. 3 show the results for the across-class design. These results are consistent with the regression analysis (discussed immediately below).

Using the regression model from Eq. (2), but excluding the control variables in the within-class design, we estimated an effect of 0.236 s.d. for the peer incentive treatment (Table 3, panel A, column 1). After including all the control variables, the measured effect rose slightly to 0.265 s.d. (panel A, column 2). The estimated ATE using matching was 0.244 s.d. (panel B, column 1). All these estimates were significant at the 1% level. In contrast, all the estimated effects from the individual incentive treatment (reported in the other part of Table 3) were small and did not significantly differ from zero.

Regression estimates based on the across-class design were nearly identical. The estimated impact of the peer incentive treatment was

Table 3
Regression and matching estimations of effects of peer incentive and individual incentive in within-class design (dependent var: post-test score).

	Peer incentive		Individual incentive	
	(1)	(2)	(3)	(4)
A: regression				
Treatment	0.236***	0.265***	−0.060	−0.061
	(0.065)	(0.067)	(0.083)	(0.078)
pre-test		0.424***		0.604***
		(0.057)		(0.048)
Male		−0.045*		−0.029
		(0.072)		(0.064)
Class dummies	No	Yes	No	Yes
R-squared	0.013	0.343	0.001	0.356
N	747	747	829	829
B: matching				
Treatment (ATE)	0.244***		−0.044**	
	(0.066)		(0.065)	
Exact matches	95%		93%	
N	747		829	

Note: Our samples are the bottom students in the within-class design (bottom treated students and their within-class controls). In panel B, exact matching is on pre-test within-class ranking (i.e. 40th student matched with 39th student, 38th student matched with 37th student, etc.). Figures in parenthesis are standard errors clustered at the class level.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

0.303 s.d. and was statistically significant at the 1% level (Table 4, panel A, column 1). The same estimation for the individual incentive treatment was 0.026 s.d. and was not statistically significant at the 10% level (Table 4, panel B, column 1).

5.4. Spillover effects and welfare analysis

There were four types of students in each experiment class: (1) BT students; (2) BC students; (3) Quartile 2 students; and (4) Quartile 1 or top (T) students. Only BT students were treated. We are additionally concerned with the spillover effects on untreated students in the experiment classes. In particular, there is a potential concern that the reshuffled high achieving students (T) who were assigned as benchmates to interact with low achieving students might suffer from the program. There is also a concern that the BC students might not be good within-class control students, because they might become demotivated when they figured out that some of their classmates were eligible for cash rewards that they were not allowed to compete for. We compared these four types of students from the experiment classes to similar types of students from the pure control classes using regression (2). The estimated treatment effects were reported in four columns of Table 4, following the order above.

The estimates for the BT students were already discussed in the previous subsection. The estimates for the other students, as reported in column 2–4, were all relatively small and indistinguishable from zero at the 10% level. In particular, the impact on the T students in the peer

Table 4
Regression estimations of treatment effects for students in experiment classes in across-class design (dependent var: post-test score).

	BT	BC	Q2	Q1 (T)
	(1)	(2)	(3)	(4)
A: peer incentive				
Treatment	0.303***	0.030	0.009	−0.026
	(0.077)	(0.077)	(0.074)	(0.057)
Pre-test	0.532***	0.596***	0.598***	0.677***
	(0.044)	(0.041)	(0.059)	(0.094)
Male	−0.097	−0.020	−0.115**	0.059
	(0.067)	(0.059)	(0.048)	(0.046)
Grade dummies	Yes	Yes	Yes	Yes
R-squared	0.242	0.264	0.090	0.100
N	958	963	877	748
B: individual incentive				
Treatment	0.026	0.088	0.029	0.071
	(0.093)	(0.071)	(0.077)	(0.059)
Pre-test	0.605***	0.581***	0.570***	0.791***
	(0.043)	(0.044)	(0.075)	(0.123)
Male	−0.104*	−0.043	−0.032	0.049
	(0.059)	(0.061)	(0.046)	(0.039)
Grade dummies	Yes	Yes	Yes	Yes
R-squared	0.244	0.251	0.066	0.152
N	998	1005	885	777

Note: Columns 1, 2, 3, and 4 estimates the impact of the direct treatment for the treated students, and spillover effects for their within-class controls, quartile 2, and quartile 1 students in the experiment classes respectively. Control groups are bottom-half, bottom-half, quartile 2, and quartile 1 students in the control classes respectively. Figures in parenthesis are standard errors clustered at the class level.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

incentive classes was negative, as most peer effects models would predict, but it was not statistically significant at the 10% level. Contrary to the de-motivating hypothesis, the treatment effects on the BC students in both the peer and the individual incentive classes were positive, although neither of them was statistically significant at the 10% level. In a word, our program did not systematically lower student test scores.

Spillover effects regarding the BC students show that using an entire class as the controls did not improve much upon our within-class design. This result suggests that using an entire school as the controls will most likely not improve much upon our current designs either.

The lack of spillover effects on untreated students (BC, Quartile 2 and T) lent further strength to our treatment effect estimations because untreated students in the experiment classes were effectively an extra control group. Moreover, peer effects operating through the interaction mechanism appear to be Pareto efficient. This finding is important. It is well known in the peer effects literature that there are cases where high achievers may be harmed by having low achieving peers (Sacerdote, 2011).

5.5. Are peer incentives more effective than individual incentives?

As described earlier, we chose 11 schools to host the peer incentive experiment, and another 12 comparable schools to host the individual incentive experiment. As discussed above (and according to our power calculations), these numbers are adequate for estimating the effect of a treatment because of our within-class and across-class evaluation designs. Because the effects were not estimated by comparing outcomes across schools, it is unlikely that sampling error or exogenous shocks at the school level were driving the observed treatment effect differences. Still, we must be cautious in concluding that the peer incentive treatment is *more* effective than the individual incentive treatment because of the small number of schools involved. We applied several methods to check this claim.

First, by inspecting the estimation results in Tables 3 and 4, we can quickly reject the hypothesis that the coefficient of peer incentive effects fell in the confidence interval of the coefficient of the individual incentive effects. The estimated 95% confidence intervals of these two treatment effects do not overlap under the within-class design and only barely overlap under the across-class design. We also regressed post-test scores on a treatment status dummy variable, where the dummy equaled zero if a given BT student received the individual treatment and one if a given BT student received the peer incentives treatment, as well as the other variables specified in regression (2). We obtained a positive coefficient for the treatment effect difference $\hat{d} = 0.280$, significant at the 5% level with clustered standard errors at school, grade, or class levels.

Second, we pooled data from both experiments from the within-class evaluation design and ran two regressions to evaluate whether the treatment effect difference was statistically significant. We first regressed the post-test scores on the incentive dummy (1 for all BT students, 0 for all BC students), the peer incentive dummy (1 for peer incentive BT students only, 0 otherwise) and the usual controls. After estimating this effect, only the peer incentive dummy coefficient (approximately 0.281 s.d.) was positive and significantly different from zero at the 5% level. We additionally ran a regression in which we interacted treatment status at the class level (1 for BT students, 0 for BC students) with treatment status at the school level (1 for peer incentive schools, 0 for individual incentive schools).[13] The coefficient for the interaction term was positive (0.312 s.d.) and statistically significant at the 1% level. The above regression results were robust when we used data from the across-class evaluation design and/or clustered the standard error at the school level.

Third, we followed the randomization inference method in Kremer et al. (2006) to test treatment effect differences by calculating an empirical distribution of the coefficient mentioned above. We randomly assigned 11 schools to be placebo peer incentive schools and the other 12 schools to be placebo individual incentive schools. Then, we calculated the placebo treatment effect difference by running the regression mentioned above. We repeated the above procedure 10,000 times and found that the simulated $\hat{d} > 0.280$ only 2% of the time. That makes it extremely unlikely that we observed such a treatment effect difference by chance. As a falsification test, we applied this method to pre-test scores instead of post-test scores. The empirical p-value was approximately 0.480, as expected. Applying the same method to the pooled regressions mentioned above produced similar results. Following these different analyses, we conclude that the peer incentive treatment appears to be more effective than the individual incentive treatment.

6. Pure benchmate effects

To estimate the conventional reduced-form pure peer effects for benchmates (i.e., pure benchmate effects), we collected official seating tables at the baseline survey to determine the students' original benchmates. The most popular seating arrangement was a two-student bench (50.55%), followed by a three-student bench (22.50%), a single-student bench (8.44%),[14] and then a bench with four or more students (4.59%). Approximately 13.92% of our sample did not have benchmate information because seating tables for eleven classrooms were missing.

Assortative peer group formations at the classroom or school level are well-known, which underlies much of the academic interests in peer effects. In our context, low achieving students were less likely to sit next to high achieving classmates, consistent with the findings in Carrell et al. (2011), who found that low achievers were more likely to interact with each other than with high achieving peers. Controlling for gender and class dummies, a one standard deviation lower pre-test score was associated with an approximately 0.190 standard deviation lower average benchmate pre-test score (details omitted). Standard errors were adjusted for intra-group correlation at the bench level. The coefficient was highly significant at the 1% level. Our reseating intervention thus significantly changed the classroom peer environment.[15]

One common way of estimating pure peer effects is to regress post-test scores on peer characteristics (such as average peer pre-test scores and average peer family income), controlling for the student's own characteristics (see regression (3) below). This approach is valid when variations in peer characteristics are exogenous to own characteristics, a condition clearly violated in our full sample because of the positive correlation in benchmate pre-test scores. We instead exploited two exogenous variations in benchmate composition created by student reseating in our peer incentive experiment classes.[16] Because the two types of exogenous variation come from entirely different sources, each with their own set of strengths and weaknesses, they can help us draw more reliable conclusions if the estimations turn out to be consistent with each other.

[13] We thank Esther Duflo for suggesting this approach.

[14] Most of these students came from classrooms that had a one-student-one-desk arrangement. If such a classroom happened to be assigned to host the peer incentive experiment, we required the BT and the assigned T students to sit close to each other.

[15] The average standardized test scores of the top students in our peer incentive experiment is 1.062, while that of the bottom students is −.813. There is a difference of 1.875 standard deviations, a very large gap.

[16] We nevertheless tried the above regressions using the percentage of benchmates who attended preschools, and the percentage of benchmates who repeated grades as two measures of peer quality. Whether a student attended preschools (or repeated grades) had a significantly positive (or negative) impact on her own test scores. These two measures are superior to lagged benchmate achievement (i.e. pre-test scores), which is likely to be determined simultaneously with own lagged achievement (Lavy et al., 2012). The estimated coefficients tended to be consistent with most findings in the peer effects literature, but were not statistically significant once own pre-test scores were controlled for.

Table 5
Regression estimations of the effects of randomly-assigned partner' pre-test scores on own test scores.

	Pre-test		Post-test		Post-test		Post-test	
	BT	T	BT	T	BT	T	BT	T
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A: main results								
Partner pre-test	−0.187	−0.012	0.209	0.006				
	(0.141)	(0.016)	(0.215)	(0.046)				
Male	0.031	0.021	−0.133	0.044	−0.122	0.039	−0.119	0.043
	(0.091)	(0.036)	(0.137)	(0.061)	(0.130)	(0.062)	(0.124)	(0.063)
Pre-test			0.366***	0.659***	0.364***	0.632***	0.365***	0.654***
			(0.086)	(0.136)	(0.085)	(0.132)	(0.086)	(0.139)
Partner post-test					0.221**	0.099***	0.328*	0.016
					(0.091)	(0.033)	(0.306)	(0.121)
Class dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.420	0.527	0.348	0.293	0.362	0.306	0.358	0.297
N	371	385	371	385	371	385	371	385
B: first-stage for IV						Partner post-test		
Partner pre-test							0.638***	0.381***
							(0.121)	(0.058)

Note: Our sample corresponding to odd-numbered columns is the set of bottom treated students in peer incentive classes. Our sample corresponding to even-numbered columns is the set of reshuffled top students in peer incentive classes. They have different sample sizes because of slightly different attrition rates. Figures in parenthesis are standard errors clustered at the class level.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

6.1. Random partner matching

Despite the fact that all the treated BT students in the peer incentive classes were randomly assigned a high-performing student as a benchmate, there was still enough naturally occurring variation in the pre-test scores of these reshuffled top students, possibly because of large score differences across classes and schools.[17] Therefore, we followed the college roommate effects literature (Sacerdote, 2001; Zimmerman, 2003; etc.) by estimating the following regression (regressing own test scores on partner lagged test scores) for pure peer effects:

$$Post\text{-}test_i = \alpha_0 + \alpha_1 Pre\text{-}test_{i'} + \mathbf{Y_i}\alpha_2 + \epsilon_\mathbf{i} \quad (3)$$

where $Post\text{-}test_i$ is BT student i's post-test score, $Pre\text{-}test_{i'}$ is the pre-test score of student i's randomly assigned benchmate (*Partner Pre-test*), and $\mathbf{Y}_i$ includes own pre-test scores, gender, and class dummies as control variables. The coefficient α_1 provides the reduced-form estimation of the pure benchmate effect for the BT students in peer incentive classes.

Similarly, the reshuffled top students (T) were randomly matched to BT students with varying pre-test scores.[18] We ran the same regression as in Eq. (3) on the T students to estimate pure benchmate effects for them. We reported the regression results for BT and T students separately in Table 5 because peer effects may be nonlinear.

Columns (1) and (2) in panel A of Table 5 report the results of the regular randomization check by running regression (3) but with $Pre\text{-}test_i$ as the outcome variable instead of $Post\text{-}test_{i'}$. The estimated α_1 was not statistically significant at the 10% level, suggesting that matching was random.[19]

Columns (3) and (4) in panel A of Table 5 report the results for the regression in Eq. (3). The estimated α_1 coefficients were not statistically significant. The absence of an effect suggests that randomly assigned benchmate pre-test scores did not significantly affect own academic performance in these two samples.

Instead of asking whether a partner's lagged test scores affected a student's current scores, we can also ask whether the benchmates' current scores were correlated with each other. The latter is a less-demanding way of identifying peer effects. Sacerdote (2001) did both using randomly assigned roommates at Dartmouth College, and only found significant peer effects in the latter specification. Following Sacerdote (2001), we also estimated the following regression about current test score correlations between randomly assigned benchmates.

$$Post\text{-}test_i = \alpha'_0 + \alpha'_1 Post\text{-}test_{i'} + \mathbf{Y_i}\alpha'_2 + \epsilon_\mathbf{i} \quad (4)$$

Columns (5) and (6) in panel A of Table 5 report these results. For the BT students, we found that a one standard deviation increase in the partner's post-test scores was associated with a 0.221 standard deviation increase in own post-test scores. The same association was weaker for the T students (approximately 0.099 s.d.). Both were statistically significant at the 5% level. The results suggest that students assigned to the same bench most likely exerted some influence on each other.

Our finding that a student's current test score is affected by her benchmate's current scores but not by her benchmate's lagged test scores is very close to Sacerdote's (2001) findings for Dartmouth roommates. Because the partners were assigned to share the same bench, their post-test scores were determined simultaneously. Because of this, the coefficient α_1 is subject to the reflection problem and cannot be interpreted as causal (Manski, 1993; Sacerdote, 2001). The results of regression (4) only show that there is a degree of correlation in benchmates' current outcomes (although the coefficient is difficult to interpret).[20,21]

[17] The pre-test scores of the reshuffled peers had a standard deviation of approximately 0.36 s.d., higher than the corresponding number in Duflo et al. (2011).
[18] The pre-test scores had a standard deviation of approximiately 0.99 s.d.
[19] No systematic relationship existed for three other baseline dummy variables: gender, whether the student had previously repeated grades, and whether the student attended preschools (results not shown).
[20] Following Duflo et al. (2011), we used partner pre-test scores as an instrumental variable (IV) for partner post-test scores in regression (4). The results, shown in Table 5, are consistent with the reduced-form estimations.
[21] A plausible potential policy is to match the bottom half of the students with the top half in the same classroom. Our results from random partner matching suggest that if the difference between the partners is large enough, the low achievers would benefit from such a policy regardless of the partner's test scores. However, we acknowledge that we do not know much about the potential effects if the difference between the partners is small. For example, the student with the highest pre-test scores in the bottom half may experience an entirely different treatment effect, depending upon whether she is matched with the student with the lowest or highest pre-test scores in the top half.

6.2. Exogenous benchmate reshuffling

Our second approach to estimating pure peer effects focused on the students in our peer incentive classes who "lost" a strong or weak peer because of our exogenous intervention. Reseating in our peer incentive classes took the top T students away from their original benchmates (T-bench), which caused a significant and yet exogenous decline in the T-bench's peer environment because their new benchmates would not be as high performing as their original, top-performing benchmates. Because there was no reseating in the pure control classes, the top students (T) and their original benchmates (T-bench) in the control classes remained next to each other during the evaluation test. In other words, the T-bench students in control classes could serve as a comparison group for the T-bench students in peer incentive classes. To estimate the treatment effect, we applied regression (2) to the T-bench students in both the peer incentive and control classes. Because the T-bench students in the peer incentive classes experienced a significant negative shock in their peer environment, we predicted that the coefficient β would be negative for the T-bench students.

In contrast, the reseating in the peer incentive classes took the BT students away from their original benchmates (BT-bench), which had a significant and yet exogenous effect of improving the BT-bench's peer environment. The BT-bench students in the control classes could serve as a comparison group. To estimate this effect, we applied the same regression (2) to the BT-bench students and their counterparts in the control classes. We predicted that β would be positive for BT-bench students.

Our trial was designed to make sure that the treated BT (or T) students were comparable to the bottom (or top) students in the control classes, but there was no guarantee that either group's original benchmates would be comparable between the treatment and control classes. Therefore, we first assessed the randomization assumption by running the regression in Eq. (2) on the BT-bench (or T-bench) students and their counterparts in the control classes but replacing $Post\text{-}test_i$ with $Pre\text{-}test_i$. The results for the BT-bench and T-bench were reported in columns (1) and (2) of Table 6, respectively. No coefficient was statistically significant at the 10% level. BT-bench/T-bench and their counterparts in the control classes had similar baseline test scores.

Columns (3) and (4) of Table 6 reported the treatment effect estimates for the BT-bench and T-bench students, respectively. The estimated treatment effect for BT-bench was indeed positive (0.087 s.d.), but it was not statistically significant at the 10% level. The estimated treatment effect for T-bench turned out to be positive as well, but it was close to zero and also not statistically significant at the 10% level.

In summary, as discussed in the two subsections above, our estimations (based on two different sources of exogenous shocks to benchmate composition) point to the same conclusion: there is no evidence of strong, causal, pure benchmate effects in our sample. As such, reshuffling benchmates alone is unlikely to be driving the observed peer incentive effects.

The literature on classroom peer effects sometimes identifies strong peer effects, but not always (see the reviews by Epple and Romano, 2011; Sacerdote, 2011). Our pure benchmate effects may be particularly weak because our results are based on benchmate reshuffling inside a given classroom. In contrast, the results in the literature are based on overall changes in classroom student body composition, which may be more likely to have a relatively greater influence on individual students.

Extra robustness checks (including heterogeneous analysis, student subjective evaluation of the programs, question-level evidence regarding testing efforts, plagiarism, etc.) are reported in the appendices. Our primary results are not driven by these alternative explanations.

7. Conclusion

Our peer incentive experiment in China paired high and low achieving students through benchmate reshuffling within classrooms and provided monetary incentives based on the weaker partner's test scores to motivate beneficial peer interactions. We found a robust impact of approximately 0.25 to 0.30 standard deviations in test scores on the treated low achieving students, and found no detrimental impacts on their high achieving partners or other classmates. In other words, encouraging classroom peer interactions resulted in a Pareto improvement. In the paper, we also tried to unpack this effect. We determined that this result was unlikely to be driven only by the individual desire to win rewards on the part of the weaker partners, as we found no such effects in a separate individual incentive experiment. We also do not believe that the results were driven by pure peer effects associated with benchmate reshuffling. The pure peer effects did not appear to be strong in our context. Given these two results, we cautiously conclude that our peer incentive treatment worked because group incentives made peer effects more effective than they would have been

Table 6
Regression estimations of the treatments effects of removing BT and T benchmates on BT-bench and T-bench students' test scores in across-class design.

	Pre-test		Post-test	
	BT-bench	T-bench	BT-bench	T-bench
	(1)	(2)	(3)	(4)
Treatment	0.141	−0.021	0.087	0.028
	(0.106)	(0.095)	(0.063)	(0.068)
Male	−0.076	−0.101*	0.009**	−0.002
	(0.076)	(0.087)	(0.063)	(0.070)
Pre-test			0.582***	0.500***
			(0.033)	(0.048)
Grade dummies	Yes	Yes	Yes	Yes
R-squared	0.011	0.009	0.413	0.252
N	672	565	672	565

Note: BT-bench and T-bench refers to original benchmates of BT and T students in peer incentive classes, respectively. They have different sample sizes mainly because of different attrition rates. Figures in parenthesis are standard errors clustered at the class level.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

otherwise. Of course, there is no definitive identification of the exact mechanism at work in the treatment effect. Specifically, we still do not know (neither in this paper nor in most other places in the literature), whether group incentives enhanced peer effects because the high achievers spent more time helping their low achieving benchmates, because the low achievers worked harder so as not to "let down the team," because the benchmate pairs spent more time studying individually instead of playing or fighting with each other, or because some other mechanisms were at play. It is well known that many channels potentially exist for peer influence (e.g. Sacerdote, 2011). Adding incentives only makes interpretation more complicated. We leave these for future research.

In terms of policy, one of the lessons of this study is that to increase desirable peer effects, policymakers must pay more attention to student motivations. Without stronger incentives, peers may not interact as much with each other as policymakers may intend. School integration policies, such as detracking, might be made more effective by paying attention to the nature of peer interactions. Our results additionally suggest that simple and inexpensive interventions exist for policymakers (or donors) who are willing to pay students to achieve better test scores. These policymakers may generate better outcomes simply by changing the functional form of the incentive contract to tap into local peer resources.

The size of our peer incentive treatment effect was similar to that of a large-scale teacher incentive pay program conducted in Indian primary schools (Muralidharan and Sundararaman, 2011). The nominal cost per student of their program was approximately 2.0 U.S. dollars (our calculation), while the corresponding measure in our program was approximately 5.7 U.S. dollars. The cost was higher in our program, but China is a wealthier country, with a nominal GDP per capita approximately 3.2 times that of India in 2010. This simple comparison shows that paying high and low achieving peers to learn as a group appears to be at least as cost-effective as offering incentives to teachers. However, we acknowledge that a more appropriate comparison should take into account the internal rate of return.

In addition to those mentioned above, this study points to several other promising directions for future research. In this paper we only examined the effects of cash incentives, but other types of rewards, whether extrinsic or intrinsic, may work more effectively to encourage peer interactions. Our experiment used an incentive contract based on score improvement. However, we cannot definitively say that this value-added contractual form should be maintained if this program were to be expanded and repeated. In this type of longer-run setting, students might very well begin to game the program, although we recognize that the pre-score could be the previous year's post-score, which was also part of a reward system and so is less likely to be gamed. If official tests, such as the high school exit exams popularly used in the pay-for-grades literature, can be used to measure student performance, our program could be expanded using a standard linear contract based on raw scores as well. This said, the value-added approach has been less-thoroughly studied than other approaches and thus deserves more research in the future. We also recognize that a tournament structure for incentives within a classroom will lead to objections by some parents, due to the competitive nature and potential negative impacts on classroom cohesion in the short- or long-run, a point that is not addressed in our current study. Finally, researchers who are interested in opening up the black box of classroom peer effects may also want to examine well-defined peer groups *inside* of classrooms. The current study only touches upon the relationship between a pair of students with very diverse performance. In the current study we do not learn what would happen if low-performing students were matched with low-performing students and faced group incentives (e.g., compared to high and high groups). This gap makes it difficult to compare the current study with tracking, an important related policy.

Appendix A. Baseline variables and balance checks at school level

From the baseline student survey, we coded 33 pre-treatment variables related to student and family characteristics (details in Section F.1). From the baseline test, we coded the following two variables:

math: the standardized math scores in the baseline test

Chinese: the standardized Chinese scores in the baseline test.

In total we have 35 baseline variables at the student level.

The quality of our survey and test appears to be good. We have two measures of student academic performance: the standardized test scores from our baseline exams (by math and Chinese), and the self-reported grades for math, Chinese, and English from the final exam of the last semester. The correlations between self-reported math (or Chinese) grades and our baseline math (or Chinese) test scores is 0.35 (or 0.22), while the correlation between our baseline and evaluation math (or Chinese) test scores is 0.55 (or 0.39). Because our baseline and evaluation tests were designed, implemented and graded using a uniform standard, it is not surprising that the correlation between our two tests is higher than the correlation between self-reported grades and our baseline math test scores. This said, the latter correlation is still quite high, which helps demonstrate the reliability of our test scores.

We also regress test scores on grades, controlling for gender and class dummies. We find the correlation to be highly significant. The results are reported in Table G.1 by math and Chinese. The self-reported grades have many missing observations. Our main analysis is based on the standardized test scores.

After randomization, 13 baseline variables were balanced across two groups of schools (12 individual incentive schools vs. 11 peer incentive schools). These 13 baseline variables are: math, Chinese, gender, birth year, repeatgrade, fatherout, motherout, elderbro, broedu, eldersis, sisedu, and selfesteem responses 2 and 10.

We adopted the rerandomization methods to conduct the random assignment at the school level. We took a random draw of 11 schools, examined the difference in means for these 13 baseline variables, and then rerandomized if the "equal mean" null hypothesis was rejected for at least one student-level variable at the 15% statistical level in a standard two-sample t-test. We repeated the process using a loop in stata until our two groups of schools were balanced in the way defined above (details available upon request).

To check whether our randomization achieved sample balance at the school level, we construct two sets of school-level variables. First, we averaged our 35 student-level variables at the school level. In our notation we put "(mean)" before the student-level variable to denote the corresponding school-level variable. For example, gender is a student-level variable, while (mean) gender is a school-level variable. Second, we coded 4 additional pre-treatment variables at the school level that are not based on student survey: classN, landline, distance, and mac.

classN: the average class size

landline: whether the school had a land line (phone connection) or not, 1: yes, 0: no.

distance: the shortest public transportation distance of the school to Tiananmeng Square, the city center (unit: kilometer, information based on maps.baidu.com)

mac: the walking distance of the school to the nearest McDonald's restaurant (unit: kilometer, information based on maps.baidu.com).

In total, we have 39 pre-treatment variables at the school levels. We report the two sample t-test for these 39 variables in Table G.2. Except for two variables, all of the 39 variables are balanced (t-test at the 5% level).[22]

[22] $2/39 \approx 5.1\%$, the probability of type-I error.

To mitigate the concerns that standard test of comparison of means might be insufficient. We also conducted two-sample Kolmogorov–Smirnov test for equality of distribution functions for these 39 variables. Only two variables failed to pass the test at the 5% level (results available upon request). So we conclude that our randomization achieved balance at the school level.

Appendix B. Implementation details

Our enumerators returned to the 23 schools in the middle of September and implemented the experiments after we completed the school, class, and individual random assignment on computer. The teams summoned the selected participating students in each experiment class (i.e., BT and T students in peer incentive classes, and BT students in individual incentive classes) to the headmaster's office and announced the program to them in the presence of their teachers and headmasters. The purpose of using such a formal setting was two-fold. First, we wanted to contact only participating students from the experiment classes and to avoid contact with the other students. Second, we wanted to make our offer appear credible. The treatment was described as a competitive scholarship program being conducted by a renowned government research institute that was aimed at boosting student academic performance.

Using a predetermined, standardized text read by the leader of each enumeration team, we told the assembled groups that because of funding and logistical limitations we could only select some students to participate. The BT students were encouraged to challenge themselves to improve as much as possible by the next test scheduled at the end of the semester. Enumerators avoided labeling the BT students as underperforming students needing particular help. Specifically, the BT students were not told that they were from the bottom half of the class. In the peer incentive classes, our enumerators ensured that the teachers made the necessary seat assignment changes according to a list we provided so that the BT students and their assigned T student partners would sit next to each other.

At the end of the meeting, we gave students an official letter describing our motivation and the key points of the program (see Section F.2 for translated letter). The explanation fit on one piece of paper. Students were required to have their parents sign and return the letters if they wanted their children to participate in our program. Students who did not want to participate were permitted to exit. Only three of the students (and/or their parents) declined our offer.

In the middle of the program, half of the treated students, who either received individual or peer incentive treatment, were randomly assigned to receive an extra parental phone call and text message intervention. The purpose of having an extra round of communication was to briefly remind the targeted parents the information they already received. No new information was provided in the extra communication stage.

In the middle of November, our enumerator team called all of the parents of the treatment students who were assigned to the communication group. The enumerators read a pre-determined standard message to parents. The content of the message essentially repeated what was in the official letter. In addition, two weeks before the evaluation test, we sent them a short text message reminding them of the approaching evaluation test for the rewards.

We define *Treatment Basic* as a binary dummy variable, which equals 1 if a student is assigned to receive individual/peer incentive treatment but not the extra parental communication treatment, and 0 otherwise. We define *Treatment Basic + Call* as a binary dummy variable, which equals 1 if a student is assigned to receive individual/peer incentive treatment and the extra parental communication treatment, and 0 otherwise. We specify the following regression:

$$Post\text{-}test_i = \alpha' + \beta_{j1} TreatBasic_i^j + \beta_{j2} TreatBasic\_Call_i^j + \gamma' \mathbf{X_i} + \epsilon_i \quad (1)$$

where $j = 1$ for individual incentive evaluation samples, and $j = 2$ for peer incentive evaluation samples. *TreatBasic*, and *TreatBasic Call* are abbreviations for the variable *Treatment Basic*, and *Treatment Basic + Call*, respectively. **Xi** include gender and class (or grade) dummies.[23]

Only approximately 70% of the students provided valid phone numbers for their parents in the pre-test survey. We can show that phone numbers are missing at random with respect to our treatment assignment (results available upon request). We dropped students with missing phone numbers to improve estimation efficiencies. The resulting power is relatively low. We only had 371 students who received the peer incentive treatment. Among them, 187 were randomly assigned to receive an extra parental communication intervention. Because of missing or wrong phone information, only 117 actually received phone calls from us.

We can use β_{j2}–β_{j1} as a point estimate of the marginal effect of having an extra round of communication with parents. The results are reported in Table G.3. The estimated basic peer incentive treatment effect in the within-class design is 0.314 (column 1, row 1), whereas the overall effect of basic peer incentive treatment plus extra communication had an estimated effect of 0.198 (column 1, row 3) only. The difference is −0.116, meaning that extra parental communication reduced peer incentive effect by 0.116 standard deviations. While both 0.314 and 0.198 were statistically significant, the differences between them is clearly not, as the estimated standard errors of these two coefficients were 0.109 and 0.087 respectively. The same results are obtained from the across-class design (column 2). Neither treatment effects from the individual incentive experiments are statistically significant (columns 3 and 4).

If we directly compare treated students who were not assigned to receive parental intervention with treated students who were assigned to receive parental intervention using regression or t-test, the estimated intention-to-treat marginal communication effects are never statistically significant (results available upon request).

Appendix C. Heterogeneous effects

Using the within-class design (shown in Table G.4, panel A), the peer incentive effects appeared to be higher for math (0.283 s.d.) than for Chinese (0.168 s.d) (columns 1 and 2). However, there is no strong evidence that the observed heterogeneous effects were statistically significant at the 5% level. Comparing the differences between math and Chinese post-test scores across control and treatment groups in a difference-in-differences-type regression, or interacting treatment dummy with gender dummy in a pooled regression failed to produce a statistically significant coefficient for the interaction terms (results available upon request). The peer incentive effects additionally appeared to be higher for girls (0.362 s.d.) than for boys (0.227 s.d) (columns 3 and 4). But the difference does not appear to be statistically significant either. The peer incentive effects did not appear to differ much by grade (grades 3 and 4 vs. grades 5 and 6, columns 5 and 6) or by pre-test scores (low vs. high, columns 7 and 8). Estimations using across-class design (shown in panel B) were consistent with the patterns discussed above.

We also would like to test whether treatment effects differ by different pair combinations. We created a partner-gender dummy that records the partner's gender for treated students in our peer incentive experiment, and a same-sex dummy that equals 1 if two partners in the peer incentive experiment have the same gender, and equals 0 otherwise.

We tried the following two regressions on the sample of treated peer incentive students. First, we regressed the evaluation test scores on own gender dummy, partner–gender dummy, and their interactions. Second,

[23] The estimation procedure follows Kremer M, Miguel E, Mullainathan S, Null C, Zwane A. 2009. "Making water safe: Price, persuasion, peers, promoters, or product design?" Work. Pap., Harvard University, Cambridge, MA.

we regressed the evaluation test scores on the same-sex dummy. In both regressions, these dummies are not statistically significant (results available upon request).

Given the limited sample size, it is not surprising that treatment effects do not differ by individual or pair characteristics. However, our heterogenous analysis does provide some extra robustness checks for our main results.

Appendix D. Rule knowledge and subjective evaluation

To be able to claim our treatments have made a difference in test scores, students must know the rules of the tournaments well. Responses from a random sample of students (N = 274) at the end of the program show that the majority of the participating students knew the prize structure well. Only 7 of 60 reshuffled high achieving students (12%), 8 of 60 treated low achieving students in peer incentive classes (13%), and 12 of 48 treated low achieving students in individual incentive classes (26%) responded that they did not know it well. On the other hand, the majority of the non-participating students did not know the prize structure well, with 57 of 106 non-participating students (54%) providing a negative response.

In addition, students' subjective evaluation of our program is a helpful metric by which to judge the plausibility of our findings. We found that students with more knowledge of the prize structure were far more likely to agree that the program made a positive impact. Moreover, students more deeply involved in the program (e.g., participating students in the peer incentive classes were more involved than participating students in the individual incentive classes) are more likely to agree that the program had a positive impact. The percentage of students in each group agreed the program was beneficial increases according to their involvement: non-participating students (38%), treated low achieving students in the individual incentive classes (64%), treated low achieving students in the peer incentive classes (70%), and reshuffled high achieving students in the peer incentive classes (83%). This order is compatible with our estimation results. Students who played a helper's role may believe they made a bigger positive impact compared to those who received help. Students who received help thought positively about our program, suggesting that they might not have been coerced into studying hard to win monetary rewards for the team.

Appendix E. Question-level evidence

To mitigate the concerns that our treatment effects could be driven by testing behaviors, we demonstrate the following two results based on question-level evidence. First, control students did not put significantly less efforts into answering the questions in evaluation tests. Second, peer incentive did not induce students to engage in more plagiarism.

First, we calculated the blank answers each student left as a percentage of the total number of questions by math and Chinese. We call this variable *Blank Answer Rate*. We conducted a two-sample t-test of this variable between the control and treatment groups by math and Chinese for both baseline (panel A, Table G.5) and evaluation (panel B, Table G.5) tests. The differences are all small, and none of them is significant at the 10% level. There is no evidence that control students take the exam less seriously in the evaluation test.

Our results for the individual incentive experiment also suggest that treated students did not try harder to win the money, but were not able to improve scores on their own. If they had tried, they probably would have left fewer blank answers, given the multiple choice nature of the exam. Our results for the peer incentive experiment also suggest that high achievers did not coerce their low aching partners to work hard to win the money — if coercion existed, there would have been strong pressures for the low achievers to leave less blank answers either.

Second, because the exam was administered by our enumerators and graded by computers, cheating should not have been a systematic problem. Here we present one piece of evidence that support the hypothesis that benchmate collusion did not drive our findings. We do not have to worry about other types of cheating because we did not observe a treatment effect for the individual incentive experiment, which offered incentives but not opportunities to cheat. We designed a statistical procedure to formally test the no collusion null hypothesis for each benchmate pair, by looking at how many answers were identical between them relative to students who had similar scores but did not sit next to the concerned students. We found out that the rate at which we rejected the null hypothesis was similar across control and treatment groups. Therefore we conclude that there was no evidence of increased collusion in the treatment group (details available upon request).

Appendix F. Raw documents

F.1. Baseline survey questions

We record below the translated baseline survey questions and the corresponding variable names in our pre-treatment balance checks and empirical analysis.

gender: 1 for boy, 2 for girl.
birth year: year of birth.
birth month: month of birth.
kindergarten: number of years of kindergarten, 1: never, 2: 1 year, 3: 2 years, 4: 3 years or above.
preschool: number of years of preschool, 1: never, 2: 1 year, 3: 2 years, 4: 3 years or above.
repeatgrade: have you repeated grades before? 1: yes, 2: no.
fatherout: where does your dad live in this semester? 1: home, 2: workplace.
motherout: where does your dad live in this semester? 1: home, 2: workplace.
livewith: whom do you live with in this semester? 1: dad, 2: mom, 3: both dad and mom, 4: grandparent, 5: grandparent-in-law, 6: others.
timeinbj year: the number of years living in Beijing.
timeinbj month: the number of months living in Beijing.
elderbro: how many elder brothers do you have?
broedu: is any of your elder brother in high school or college (or have competed any one of them)? 1: yes, 2: no.
eldersis: how many elder sisters do you have?
sisedu: is any of your elder sister in high school or college (or have competed any one of them)? 1: yes, 2: no.
youngerbro: how many younger brothers do you have?
youngersis: how many younger sisters do you have?
transport: how do you come to school? 1: walk, 2: bus, 3: school shuttle, 4: bicycle, 5: others.
transport time: how long does it take for you to come to school? 1: less than 15 min, 2: 15 to 30 min, 3: 30 min to 1 h, 4: more than 1 h.
final chinese: what is your grade in the most recent final exam on Chinese? (0–100).
final math: what is your grade in the most recent final exam on math? (0–100).
final english: what is your grade in the most recent final exam on English? (0–100).
purchase: what did your family buy for you in the past semester? 1: study.

materials, 2: books that are not required, 3: cd, 4: game station, 5: E-learning.
resources, 6: computer. Students can choose multiple items. For convenience of statistical computation we only focus on the first item answered by the students.
selfesteem 1 to selfesteem 10: 10 self esteem questions. 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree.

F.2. Letters to parents

The following is the translated letter to the parents whose children were invited to participate as the treated students in the peer incentive experiment.

With financial support from a charity fund, the Center for Chinese Agricultural Policy in Chinese Academy of Sciences is conducting a "Challenge Yourself" scholarship pilot project in the school of your child. Your child (name of the child) has been granted an opportunity to participate in the program. We have arranged for a high-performing classmate to serve as a new benchmate for your child. Only approximately 10 students in a class have been invited to participate.Your child already took a short, standardized exam in early September. At the end of the semester, your child will be asked to take another short, standardized exam. The three students from all the participants in each class who improve the most will be given a reward of 100Y, 50Y and 50Y, respectively. If your child wins a reward, his/her benchmate will win another reward of the same amount.For example, if you child wins 50Y, his/her benchmate will get another 50Y. The rewards, together with certificates from the school, will be distributed in an open ceremony.Please indicate whether you want your child to participate in the program, sign your name below, and return the following section to the school.

1. Yes. I allow my child to participate in the program.

2. No. I do not allow my child to participate in the program.

Signature:

The letter to the parents whose children were invited to participate in the peer incentive experiment as helpers has the role of the children reversed. The letter to the parents whose children were invited to participate in the individual incentive experiment is identical but without any reference to benchmates.

Appendix G. Tables

Table G.1
Regression evidence of high correlation between standardized test scores and self-reported administrative grades in the baseline period.

	Math scores	Chinese scores
	(1)	(2)
Math grades (self reported)	0.0219***,*	
	(0.0016)	
Chinese grades (self reported)		0.0161***,**
		(0.0015)
Gender dummy	Yes	Yes
Class dummies	Yes	Yes
R-squared	0.200	0.1788
N	3995	3956

Note: Figures in parenthesis are standard errors clustered at the class level. Math or Chinese grades are self-reported grades in the most recent final exam. It has a scale of 0–100.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table G.2
Balance checks between peer and individual incentive schools (two-sample t-test, *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$).

Var. name	Difference	Peer	Individual
(Mean) gender	−0.00796	1.441	1.449
	(0.0165)		
(Mean) birth year	0.0236	1998.8	1998.8
	(0.117)		
(Mean) birth month	−0.0935	6.232	6.326
	(0.116)		
(Mean) kindergarten	−0.00585	2.012	2.018
	(0.0573)		
(Mean) preschool	0.0206	2.093	2.072
	(0.0350)		
(Mean) repeatgrade	0.00749	1.639	1.631
	(0.0191)		
(Mean) fatherout	0.00286	1.167	1.164
	(0.0134)		
(Mean) motherout	0.00637	1.106	1.099
	(0.0129)		
(Mean) livewith	0.0143	3.037	3.023
	(0.0311)		
(Mean) timeinbj year	0.477**	5.351	4.873
	(0.192)		
(Mean) timeinbj month	0.335**	6.130	5.794
	(0.137)		
(Mean) elderbro	0.0268	0.423	0.396
	(0.0423)		
(Mean) broedu	−0.00890	1.828	1.837
	(0.0146)		
(Mean) eldersis	−0.00887	0.493	0.502
	(0.0278)		
(Mean) sisedu	−0.00785	1.807	1.815
	(0.0146)		
(Mean) youngerbro	0.00289	0.384	0.381
	(0.0305)		
(Mean) youngersis	0.0131	0.285	0.272
	(0.0109)		
(Mean) transport	0.194	1.983	1.789
	(0.150)		
(Mean) transport time	0.0574	1.610	1.553
	(0.0526)		
(Mean) final chinese	0.510	83.55	83.04
	(1.334)		
(Mean) final math	1.328	84.67	83.34
	(1.339)		
(Mean) final english	−0.757	75.67	76.42
	(2.729)		
(Mean) purchase	0.0495	2.914	2.865
	(0.111)		
(Mean) selfesteem 1	−0.0249	2.129	2.154
	(0.0327)		
(Mean) selfesteem 2	0.00625	2.326	2.320
	(0.0250)		
(Mean) selfesteem 3	0.0422	2.991	2.948
	(0.0372)		
(Mean) selfesteem 4	0.00334	2.099	2.096
	(0.0197)		
(Mean) selfesteem 5	−0.00943	2.205	2.214
	(0.0309)		
(mean) selfesteem 6	0.0129	2.173	2.160
	(0.0293)		
(Mean) selfesteem 7	−0.0421	2.045	2.087
	(0.0287)		
(Mean) selfesteem 8	−0.0165	1.878	1.894
	(0.0227)		
(Mean) selfesteem 9	0.0436	2.821	2.777
	(0.0371)		
(Mean) selfesteem 10	0.00565	2.554	2.549
	(0.0298)		
(Mean) math	0.0361	0.0275	−0.00852
	(0.0803)		
(Mean) chinese	0.0282	0.0148	−0.0134
	(0.0915)		
classN	0.710	42.58	41.87
	(3.359)		

(continued on next page)

Table G.2 (*continued*)

Var. name	Difference	Peer	Individual
landline	−0.121	0.545	0.667
	(0.212)		
distance	2.770	24.74	21.97
	(3.169)		
mac	1.017	4.100	3.083
	(0.656)		
N		11	12

Note: variables defined in Section F.1. (Mean) denotes aggregation at the school level.

Table G.3
Regression estimations of effects of basic treatment and additional parental communication treatment.

	Outcome variable = post-test			
	Peer incentive		Individual incentive	
	(1)	(2)	(3)	(4)
	Within class	Across class	Within class	Across class
Treatment BASIC	0.314***	0.345***	−0.102*	0.041
	(0.109)	(0.112)	(0.089)	(0.116)
Treatment basic + call	0.198**	0.279**	−0.047	0.066
	(0.087)	(0.108)	(0.096)	(0.108)
Pre-test	0.426***	0.509***	0.575***	0.628***
	(0.066)	(0.055)	(0.050)	(0.061)
Male	0.021	−0.030	−0.031	−0.076
	(0.089)	(0.081)	(0.087)	(0.074)
Grade dummy	No	Yes	No	Yes
Class dummy	Yes	No	Yes	No
R-squared	0.350	0.224	0.360	0.261
N	522	645	546	658

Note: Our samples are restricted to the bottom half students with non-missing phone numbers in the within-class design (column 1 and 3) and across-class design (column 2 and 4). Figures in parenthesis are standard errors clustered at the class level.
* p < 0.1.
** p < 0.05.
*** p < 0.01.

Table G.5
Comparison of blank answer rate in the within-class design.

	Peer incentive			Individual incentive		
	(1)	(2)	(3)	(4)	(5)	(6)
	Diff.	Control	Tr.	Diff.	Control	Tr.
A: baseline						
Math	0.00119	0.0235	0.0223	0.00447	0.0315	0.0270
	(0.00568)			(0.00605)		
Chinese	−0.00586	0.166	0.172	−0.00528	0.175	0.180
	(0.0202)			(0.0187)		
B: evaluation						
Math	0.00434	0.0215	0.0171	−0.00625	0.0199	0.0261
	(0.00599)			(0.00613)		
Chinese	0.0125	0.0691	0.0565	−0.0144	0.0618	0.0762
	(0.0117)			(0.0118)		
N		376	371		418	411

Note: "Diff" refers to the difference between control and treatment groups in a t-test. *p < 0.1, **p < 0.05, ***p < 0.01.

References

Angrist, Joshua, Lavy, Victor, 2009. The effects of high stakes high school achievement awards: evidence from a randomized trial. Am. Econ. Rev. 99 (4), 1384–1414.

Angrist, Joshua, Lang, Daniel, Oreopoulos, Philip, 2009. Incentives and services for college achievement: evidence from a randomized trial. Am. Econ. J. Appl. Econ. 1 (1), 136–163.

Babcock, Philip, Hartman, John, 2011. Coordination and contagion: friendship networks and peer mechanisms in a randomized field experiment. UCSB Department of Economics working paper.

Babcock, Philip, Bedard, Kelly, Charness, Gary, Hartman, John, Royer, Heather, 2010. Letting down the team? Evidence of social effects of team incentives. UCSB Department of Economics working paper.

Bettinger, Eric P., 2012. Paying to learn: the effect of financial incentives on elementary school test scores. Rev. Econ. Stat. 94 (3), 686–698.

Blimpo, Moussa P., 2010. Team Incentives for education in developing countries: a randomized field experiment in Benin. Working paper.

Boning, Brent, Ichniowski, Casey, Shaw, Kathryn, 2007. Opportunity counts: teams and the effectiveness of production incentives. J. Labor Econ. 25 (4), 613–650.

Brock, William A., Durlauf, Steven N., 2001. Interactions-based models. In: Heckman, James, Leamer, Edward (Eds.), Handbook of Econometrics, vol. 5. Elsevier, Amsterdam, p. 3297C3380.

Bruhn, Miriam, McKenzie, David, 2009. In pursuit of balance: randomization in practice in development field experiments. Am. Econ. J. Appl. Econ. 1 (4), 200–232.

Table G.4
Regression estimations of treatment effects by groups (dependent var: post-test score).

	Subject		Gender		Grade		Pre-Test	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Math	Chinese	Male	Female	3–4	5–6	Low	High
A: within-class design								
Peer inctv.	0.283***	0.168**	0.227**	0.362***	0.249**	0.291***	0.283**	0.259***
	(0.077)	(0.079)	(0.108)	(0.116)	(0.117)	(0.077)	(0.111)	(0.070)
R-squared	0.316	0.218	0.390	0.378	0.326	0.357	0.216	0.377
N	746	742	457	290	371	376	373	374
Individual inctv.	0.021	−0.120	−0.102	−0.014	−0.043	−0.076	−0.072	−0.041
	(0.071)	(0.088)	(0.107)	(0.104)	(0.125)	(0.095)	(0.118)	(0.097)
R-squared	0.282	0.263	0.394	0.378	0.352	0.372	0.334	0.248
N	826	823	488	341	431	398	412	417
B: across-class design								
Peer inctv.	0.329***	0.202**	0.287***	0.318***	0.153*	0.437***	0.321***	0.237***
	(0.079)	(0.084)	(0.104)	(0.103)	(0.123)	(0.093)	(0.104)	(0.089)
R-squared	0.218	0.137	0.230	0.270	0.226	0.285	0.124	0.188
N	956	954	561	397	482	476	478	480
Individual inctv.	0.104	−0.047	−0.000	0.067	−0.042	0.099	−0.022	0.078
	(0.088)	(0.096)	(0.108)	(0.108)	(0.142)	(0.119)	(0.112)	(0.112)
R-squared	0.204	0.161	0.233	0.266	0.243	0.266	0.191	0.085
N	994	993	581	417	511	487	499	499

Note: Pre-test "Low" (or "High") in column 7 (or 8) are defined as those students with pre-test scores in the bottom (or top) half of the distribution. Control variables in OLS regressions include *Male*, *pre-test*, and class dummies (replaced with grade dummies in the across-class design). Figures in parenthesis are standard errors clustered at the class level.
* p < 0.1.
** p < 0.05.
*** p < 0.01.

Burke, Mary A., Sass, Tim R., 2011. Classroom peer effects and student achievement. FRB Boston Public Policy Discussion Papers Series, paper no. 11–5.
Carrell, Scott E., Hoekstra, Mark L., 2010. Externalities in the classroom: how children exposed to domestic violence affect everyone's kids. Am. Econ. J. Appl. Econ. 2 (1), 211–228 (18).
Carrell, Scott E., Fullerton, Richard L., West, James E., 2009. Does your cohort matter? Measuring peer effects in college achievement. J. Labor Econ. 27 (3), 439–464.
Carrell, Scott E., Sacerdote, Bruce, West, James E., 2011. From natural variation to optimal policy? The Lucas critique meets peer effects. No 16865 NBER Working Papers. National Bureau of Economic Research.
Chan, Tat Y., Li, Jia, Pierce1, Lamar, 2010. Compensation and peer effects in competing sales teams. Olin Business School working paper, Washington University in St. Louis.
Cooley, Jane, 2009. Desegregation and the achievement gap: do diverse peers help? Working Paper.
Ding, W., Lehrer, S., 2007. Do peers affect student achievement in China's secondary schools? Rev. Econ. Stat. 89, 300–312.
Duflo, Esther, Dupas, Pascaline, Kremer, Michael, 2011. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. Am. Econ. Rev. 101 (5), 1739–1774.
Echenique, Federico, Fryer, Roland, 2007. A measure of segregation based on social interactions. Q. J. Econ. 122 (2).
Epple, Dennis, Romano, Richard, 2011. Peer effects in education: survey of the theory and evidence. Handbook of Social Economics, vol. 1B. Elsevier, pp. 1053–1163.
Figlio, David N., 2007. Boys named Sue: disruptive children and their peers. Educ. Finance Policy 2 (4), 376–394.
Foster, G., 2006. It's not your peers, and it's not your friends: some progress toward understanding the educational peer effect mechanism. J. Public Econ. 90 (8–9), 1455–1475.
Fryer Jr., Roland G., 2011. Financial incentives and student achievement: evidence from randomized trials. Q. J. Econ. 126 (4), 1755–1798.
Gibbons, Stephen, Telhaj, Shqiponja, 2011. Peer effects: evidence from secondary school transition in England. Working Paper.
Giorgi, G.D., Pellizzari, M., Redaelli, S., 2009. Be as careful of the company you keep as of the books you read: peer effects in education and on the labor market. NBER Working Paper 14948.
Hamilton, Barton H., Nickerson, Jack A., Owan, Hideo, 2003. Team incentives and worker heterogeneity: an empirical analysis of the impact of teams on productivity and participation. J. Polit. Econ. 111 (3), 465–497.
Imberman, Scott, Kugler, Adriana, Sacerdote, Bruce, 2012. Katrina's children: evidence on the structure of peer effects. Am. Econ. Rev. 102 (5), 2048–2082.
Jackson, C. Kirabo, 2010. A little now for a lot later: a look at a Texas advanced placement incentive program. J. Hum. Resour. 45 (3), 591–639.
Johnson, D.W., Johnson, F., 1997. Joining together: group theory and group skills, 6th ed. Allyn & Bacon.
Kinsler, Josh, 2010. School discipline: a source or salve for the racial achievement gap? Working Paper.
Kremer, Michael, Bloom, Erik, King, Elizabeth, Bhushan, Indu, Clingingsmith, David, Loevinsohn, Benjamin, Hong, Rathavuth, Brad Schwartz, J., 2006. Contracting for health: evidence from Cambodia. Working paper.
Kremer, Michael, Miguel, Edward, Thornton, Rebecca, 2009. Incentives to learn. Rev. Econ. Stat. 91 (3), 437–456.
Lavy, Victor, Schlosser, Analia, 2011. Mechanisms and impacts of gender peer effects at school. Am. Econ. J. Appl. Econ. 3 (2), 1–33.
Lavy, Victor, Passerman, D., Schlosser, A., 2012. Inside the black box of ability peer effects: evidence from variation in low achievers in the classroom. Econ. J. 122 (559), 208–237.
Lyle, David S., 2007. Estimating and interpreting peer and role model effects from randomly assigned social groups at West Point. Rev. Econ. Stat. 89 (2), 289–299.
Manski, Charles F., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60 (3), 531–542.
Mauldon, J., Malvin, J., Stiles, J., Nicosia, N., Seto, E., 2000. The impact of California's Cal-Learn demonstration project, final report. UC Data Archive and Technical Assistance, UC Data Reports: Paper CLFE.
Muralidharan, Karthik, Sundararaman, Venkatesh, 2011. Teacher performance pay: experimental evidence from India. J. Polit. Econ. 119 (1), 39–77.
Podgursky, Michael J., Springer, Matthew G., 2007. Teacher performance pay: a review. J. Policy Anal. Manage. 26 (4), 909–949.
Rawlings, Laura B., Rubio, Gloria M., 2005. Evaluating the impact of conditional cash transfer programs. World Bank Res. Obs. 20 (1), 29–55.
Sacerdote, Bruce, 2001. Peer effects with random assignment: results for dartmouth roommates. Q. J. Econ. 116 (2), 681–704.
Sacerdote, Bruce, 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 EducationElsevier.
Slavin, Robert.E., 2010. Can financial incentives enhance educational outcomes? Evidence from international experiments. Educ. Res. Rev. 5 (1), 68–80.
Spencer, M.B., Noll, E., Cassidy, E., 2005. Monetary incentives in support of academic achievement: results of a randomized field trial involving high-achievement, low-resource, ethnically diverse urban adolescences. Eval. Rev. 29, 199–222.
Stinebrickner, Ralph, Stinebrickner, Todd R., 2006. What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. J. Public Econ. 90 (8–9), 1435–1454.
Stinebrickner, Ralph, Stinebrickner, Todd R., 2008. The causal effect of studying on academic performance. B.E. J. Econ. Anal. Policy 8 (1), 1–53 ((Frontiers), Article 14).
Zimmerman, D.J., 2003. Peer effects in academic outcomes: evidence from a natural experiment. Rev. Econ. Stat. 85 (1), 9–23.