The Developing Economies IDE-JETRO

The Developing Economies 50, no. 4 (December 2012): 334–50

REMEDIAL EDUCATION WITH INCENTIVIZED PEER-TUTORING: EVIDENCE FROM MIGRANT CHILDREN SCHOOLS IN CHINA

Tao LI

Faculty of Social Sciences and Humanities, University of Macau, China

First version received October 2011; final version accepted June 2012

This paper evaluates the impact of a remedial education program performed in two Chinese migrant children schools. The program used a conditional cash transfer contract to encourage peer tutors to help their underperforming classmates. Incentivized peer-tutoring significantly improved the academic performance of the underperforming tutees over a semester. The impact persisted in the second semester when the intervention was removed. Our results suggest that nongovernment organizations and government agencies can increase their capacity of remedying education by incorporating incentivized peer-tutoring into their existing merit-based scholarship and conditional cash transfer programs.

Keywords: Conditional cash transfer; Peer-tutoring; Remedial education
JEL classification: O11, I21

I. INTRODUCTION

Government and nongovernment agencies are increasingly using conditional cash transfer programs to reduce the negative effects of poverty. In the field of education alone, many such programs have been implemented around the globe in the last decade (see a review in Slavin 2009). However, relying solely on cash incentives to remedy education also seems logically unrealistic; in this light, a more integrated intervention strategy could be more effective. This paper evaluates such a program that encompasses elements from both conditional cash transfer and peer-assisted learning.

Peers constitute an important educational resource. Even though peer-tutoring is different from formal teaching, their functions are similar. Extensive research in educational psychology, which has mainly been conducted in industrialized countries, has demonstrated that peer-assisted learning is a very effective teaching method (Topping 2005). However, hardly any research has been done in developing countries, where peer resources are arguably more important because other resources are in serious shortage and many students do not learn effectively (Glewwe and Kremer 2006). Even in the poorest places on earth, a classroom

doi: 10.1111/j.1746-1049.2012.00178.x

lacking everything from a blackboard to trained teachers is almost certain to enjoy the presence of some smart students. Making better use of these “wild” human resources provides one solution.

Many teachers in China have tried to implement peer-assisted learning in their classrooms (e.g., Wang 2006; Xie 2008; Zhao 2009). However, it is typically done in an *ad hoc* and spontaneous fashion. What usually happens is that some teachers may request a high-performing student to help another under-performing student in the same classroom. These teachers are typically neither trained in the method nor have the time or incentive to provide extra supervision—a problem more serious in underdeveloped rural areas. The tutors and tutees are left to sort things out by themselves, which is obviously not easy. Because of these limitations, peer-assisted learning remains at a rudimentary level in China.

The experiences of the industrialized countries suggest that successful peer-assisted learning is a very skill-intensive and time-consuming teaching method (Topping 2000). Trained teachers need to play a central role. Tutoring protocols need to be carefully designed. Regular tutoring sessions need to be organized and supervised by the teachers. These crucial ingredients of the traditional peer-tutoring program are typically missing in most Chinese schools.

The purpose of this paper is to test the effectiveness of a peer-tutoring program specifically designed for developing countries. If students respond to economic incentives as adults do, a properly designed labor contract for peer tutors may boost their effort level and improve tutoring efficiency in the absence of extra teacher interventions. This motivated us to add an incentive contract on top of a streamlined peer-tutoring program. Tutors (strong students) in our one-to-one tutoring programs were eligible to compete for a prize in a tournament. Both the chance of getting the prize and its size were conditional on the test score improvement of the tutee (weak student) who was paired with the tutor. To simulate the realities in developing countries, we required neither structured peer interactions nor teacher monitoring.

To test our idea, we ran a field experiment in two migrant children primary schools in Shanghai, China. Such schools educate millions of poor migrant children (see Section II). With the prize of an official certificate the program showed improvements in the grades of under-performing children by 0.615 standard deviations over a semester. Adding a monetary reward (labeled a scholarship in our program) to the prize made the impact larger and more robust. Even though there is evidence of heterogeneity in program effects, in most program classes the estimated impact was about 0.9 standard deviations over a semester. A large part of the program impact persisted in the second semester when the intervention was removed.

Our program effects are consistent with the findings in the psychology literature on traditional peer-tutoring programs. The median range of the program effects for the tutees reviewed in Cohen, Kulik, and Kulik (1982) is between 0.5 standard deviations and 0.8 standard deviations (see also Robinson, Schofield, and Steers-Wentzell 2005).

As the focus of research gradually shifts from enrolment toward education quality, researchers are increasingly interested in remedial education programs in developing countries (Glewwe and Kremer 2006). According to Banerjee *et al.* (2007) and Glewwe and Kremer (2006), it is not easy to improve education quality with traditional methods. Banerjee *et al.* (2007) report positive effects (0.27 s.d.) of a rare successful program in urban India that hired local young women (*balsakhi*) on a flat wage rate to work on basic skills with underperforming children from poor families (see also Mischo and Ludwig [2002] and Dang and Rogers [2008] for other private paid tutoring in both developed and developing countries). They suggest that inputs targeted specifically to helping weaker students learn may be effective. Our programs are designed in a similar vein; the main difference is that our programs used peer tutors and tournament-based conditional cash transfer. Our program has a larger impact and is more flexible to implement in our context.

Sections II and III describe the background and design of our programs, respectively. Section IV describes the design of the evaluation. In section V, we introduce the main program effects for the tutees. Section VI concludes the paper.

II. BACKGROUND

The majority of migrant children schools is privately owned, and is set up to serve the children of China's estimated 150 million poor migrant workers. Because they charge migrants much lower fees than local public schools, they are the most popular choice for migrants who bring their children to live with them. In Beijing alone, about 300,000 to 500,000 migrant children were studying in about 350 migrant children schools in 2009. This size is roughly equal to the size of the local students attending local public schools.[1] Across China millions of migrant children are attending such schools.

According to the results of a large-scale standardized test across China, migrant children are the weakest link in the Chinese education system (Lai *et al.* 2012). Migrant children are the poorest performing group of students in the entire country—poorer than children attending local public schools in those underdeveloped rural areas. Improving education quality for migrant children has important

[1] The number estimations for the migrant children and their schools are based on our extensive field study in Beijing. They are consistent with numbers often cited in the media (e.g., *Christian Science Monitor*, November 7, 2008). The comparison with local students is based on official statistics.

implications for China now and undoubtedly even more in the future as the trend of urbanization continues.

The responsibility of providing primary and secondary education lies with local governments in China. For local residents with *hukou* (household registration card, an entitlement similar to local citizenship), primary education is now almost free. However, migrants without *hukou* have to pay a very high fee to enter the local public schools. Most migrants cannot afford it, creating a huge demand for paid education services provided by migrant children schools, regardless of their quality. Because students are geographically very mobile, curriculum in these schools typically follows standard national guidelines. In China, the primary school curriculum is highly centralized. The Ministry of Education makes sure that textbooks are more or less similar across the nation, and provides rigid guidelines regarding every aspect of the curriculum, including subjects offered, or even how each individual lecture should be taught. The teachers are expected to closely follow these guidelines. Migrant children schools also offer standard subjects such as math, Chinese, English, physical education, music, etc. Just like other Chinese schools, they put a lot more emphasis on math, Chinese, and English, because these will eventually be tested in the prestigious national college entrance exam. Students typically spend the whole day in class, and go home only in late afternoon.

Local governments not only refuse to lower the discriminating barrier so that migrant children can enter the local public schools, but also try to control or even eliminate migrant children schools. The motive is probably control over permanent migration. In 2006 the Beijing city government began a campaign to shut down over 200 migrant children schools. The Shanghai city government followed suit, and set a goal of eliminating all migrant children schools that did not meet the high standards set by the local government by 2010. However, because the local governments have no plan to educate all migrant children with their own resources, it is impossible to completely root out these schools. When a school is closed down by the government, it usually reopens in another location with a different name.

Financial constraints and adverse government policy seriously crippled the development of migrant children schools. Despite the fact that these schools are located in China's richest cities, they resemble schools in poor rural areas that lag behind the urban public schools by several decades. Given the political realities, it is almost impossible to improve education quality in these migrant schools through more traditional methods. We are therefore left with very few options. Improving peer-assisted learning is one of them.

In our study the local government (i.e., the Shanghai local government) had established School A only one year before our experiment. The government sent a formal teacher from a local public school to serve as the headmaster; this was the only staff member on a regular government payroll.[2] School B, a more typical migrant children school, was several years old and a private school owned by its headmaster. Because of the difference in ownership, School A had better and self-owned facilities, possibly better school administration, and possibly higher teacher morale. The most important benefit, however, was government recognition. Note that migrant schools are often closed by the government at will. Since School A was recognized by the local government, it was free from such arbitrary treatment.

The teachers and students in these two schools were otherwise quite similar. The students were all migrant children. There were 419 students in School A (excluding grade 1) in 2007. The average number of students per class was 47. There were 692 students in School B (also excluding grade 1) in 2007. The average number of students per class was 53. The class size in both schools was quite big. Student tuition and fees (about RMB500 or US$60 a semester) were the main source of funding in each school. There was no government funding of any sort in either school.

The conditions in migrant children schools make it impossible to implement a traditional peer-tutoring program. For example, tutor–tutee study sessions—typical in traditional tutoring programs—are not feasible since migrant children schools have limited available space. Also, students have a full-day schedule. And it is neither safe to keep students after school nor possible to do so, since everyone wants to go home to dinner after a long school day. Students may live far apart from each other in a large city like Shanghai. This is also the reason we did not want to randomly assign tutor and tutee (see Section III). We were hoping that students living close to each other would team up and have more interaction after school, but there was no way for us to require them to do so in such a large city.

Teachers are hired on a contractual basis at a low wage (typically about US$100, or RMB 700–800 per month). Job loyalty is therefore very low. For example, a salesgirl with a senior high school diploma might come to work as a teacher for a year or two, and then move on to another job unrelated to teaching. To save money, student–teacher ratios are typically quite high. Asking these teachers to do anything beyond their normal teaching duties, such as supervising the tutoring process, is unreasonable. Additionally, schools have little incentive to improve teaching. They are for-profit businesses with a short expected time horizon. They may be closed down by the government at any time simply because the government does not want them to exist. Whether the schools provide high-quality teaching is not a concern of the government.

[2] One strategy the government takes is to retain and control a small number of migrant schools after closing down others. This is because of strong pressure from the central government, the media, and the international community. Then some limited resources may be provided to these selected schools. To reduce its payout, the local government would not provide help to the migrant schools without controlling the numbers of such schools.

While some of these scenarios are unique to China, most are common in many developing countries. In general, the conditions that have made traditional peer-tutoring programs successful in developed countries are missing in our context.

III. PROGRAM DESIGN

Consider one program class in our study. At the beginning of the academic semester, we invited an equal number of weak and strong students to participate in our Peer-Tutoring Program. Priority was given to those who did worst/best in the final exams in the previous semester.[3] Information sessions were organized in each class to introduce the program details to the prospective students. These students brought home a formal invitation letter that also listed the program outlines. We asked the parents of the invited children to sign and return the invitation letters if they wanted their children to participate.

Each weak student was matched with a strong student in his or her own class who served as a peer tutor. Both had a say in choosing their partner. The voluntary matching process was coordinated by our enumerators and the teachers. We did not opt for random assignment mainly because we believed that teachers and students could use their private information to improve the quality of matching, which may contribute to better tutoring outcomes.

We encouraged the tutors and tutees to interact more frequently between classes and at lunchtime but we did not formally require them to do anything during school or after school, nor did we ask the teachers to pay special attention to our programs. Our design was intended to reduce external involvement in the project such that its success did not depend on such efforts. As we discussed in detail in Section II, more structured peer interaction and teacher monitoring was difficult to implement in our context. According to our interviews at the end of the program, informal peer-tutoring did occur. For example, the tutors helped the tutees do their problem sets, explained to them things they did not understand well, and so on.

The key to our experiment was a formal incentive contract (tournament) designed to encourage the tutors' efforts. The output of a tutor's effort was measured by improvement in the academic performance of his or her partner over the semester. The teacher's behavior and disciplinary environment did not affect the effectiveness of the tournament because they are common shocks to all tutors in the same class (Lazear and Rosen 1981). The improvement was defined as the difference between the rank of the total scores of the semester finals immediately before and following the program.

[3] Since ranking of the final was public information in these schools our invitation did not reveal additional information.

We did not use simple linear contracts, which would have provided teachers with the perverse incentive to inflate tutees' grades. Shastry and Linden (2009) provide one example that teachers in developing countries may inflate student performance to help them get more resources from aid agencies. Tournament helped us fix the total amount of resources provided to a particular class (RMB 300 per monetary-program class in our design) so that teachers would have minimal incentives to distort the educational outcome. Using a linear contract also requires that different tests are comparable so that students could be paid according to the score difference. This is unrealistic in poor countries. Note that in China even the most important college entrance exam scores cannot be compared over different years. It is probably not a coincidence that the most well-known cash incentive program performed in a poor country used tournament as well (Kremer, Miguel, and Thornton 2009).

Our experiment provided two types of incentive contract to the peer tutors: the nonmonetary incentive peer-tutoring program (or nonmonetary program) and the monetary incentive peer-tutoring program (or monetary program). Students and parents were aware of the types of program they would participate in at the very beginning of the semester.

Let us first consider the monetary program. A typical program class in our experiment had ten peer tutors and ten weak students. The top prize of RMB 100 (about one-fifth of a semester's tuition and fees) was awarded to the tutor whose partner made the largest improvement during the semester.[4] Four second prizes of RMB 50 went to the four tutors whose partners made the next four largest improvements. Five third prizes (small gifts) were awarded to the rest of the tutors. An official certificate of award, which clearly stated the achievement and the level of award, accompanied each prize. There was also public recognition at each school's awards assembly held for students and teachers. The nonmonetary program was identical, except that monetary rewards were removed.

We were careful not to provide direct incentives to the tutees because we wanted to clearly identify the consequences of strengthening the tutors' incentives alone. It is entirely plausible that allowing the winning tutor to share the prize with her partner (a type of group incentive, which is more difficult to interpret) would improve our program effects and make the financial rewards more equitable, but this would be an entirely different economic problem. In another working paper we report very encouraging program effects from a large-scale randomized experiment using the group-incentive scheme to motivate peer-tutoring.

[4] The size of the prize is relatively large. In a day trip (which cost RMB 70 per student) organized by School B, 56.2% of the students did not go because of economic reasons.

IV. EVALUATION DESIGN

A. *Sample*

Both schools in our study had classes ranging from grade 1 to grade 6. Each grade contained one to three classes, which were usually taught by the same set of teachers. We excluded all grade 1 students.[5] This left us with ten classes from School A and 13 classes from School B. We randomly assigned ten classes (from different school/grade combinations) to participate in the monetary program, and four classes to participate in the nonmonetary program. The remaining nine classes were comparison classes, which did not receive any intervention.

The intended treatment was not carried out in two School B classes, which included one monetary program class from grade 5, and one nonmonetary program class from grade 6.[6] These two were used as additional comparison classes in our analysis. We later discovered that the historical academic information for the only School A class assigned to the nonmonetary program was missing, and thus had to drop it from our analysis. This left us with nine monetary program classes, two nonmonetary program classes (both from School B), and 11 comparison classes.

We were interested mainly in the program effects for the tutees (i.e., the treated weak students). In a typical program class with 50 students, we invited ten or fewer students with the lowest pre-treatment scores to participate in our programs. In classes with fewer (or more) than 50 students we invited fewer (or more) weak students to participate in our program so that the proportion of the invited weak students was about 20% of the class size. Table 1 sets out the basic structure of our field experiment.

An evaluation sample (including control and treatment students) was used to calculate program effects. The evaluation sample for calculating nonmonetary program effects was made up of treated weak students in two nonmonetary program classes and corresponding control students. The treated students (19 in total) were weak students from these two nonmonetary program classes (class 3.2 and class 5.1 from School B). Because of the small number of nonmonetary program classes, to reduce estimation noises, we restricted the corresponding control students to those that study in nearby comparison classes of the same school/grade combination. The control students were those weak students (31 in total) from other comparison classes (3.3, 5.2, and 5.3) in grades 3 and 5 from School B. Class 3.1 was not included because it is a monetary program class. So the evaluation sample of the nonmonetary program included 19 treated students and 31 control students (see columns 1 and 2 of Table 2).

[5] They are not fit for our program on a regular basis as they do not have previous-semester scores in the fall semester.

[6] It seems that teacher negligence was the main cause of the problem at least in one class.

TABLE 1

Evaluation Sampling Design

	No. of Classes in School A				No. of Classes in School B			
	Control	Monetary	Nonmonetary	Total	Control	Monetary	Nonmonetary	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Grade 2	1 (7)	1 (8)	0 (0)	2 (15)	1 (10)	1 (9)	0 (0)	2 (19)
Grade 3	1 (10)	1 (4)	0 (0)	2 (14)	1 (10)	1 (9)	1 (9)	3 (28)
Grade 4	1 (9)	1 (8)	0 (0)	2 (17)	1 (10)	1 (7)	0 (0)	2 (17)
Grade 5	1 (8)	1 (9)	0 (0)	2 (17)	2 (21)	0 (0)	1 (10)	3 (31)
Grade 6	0 (0)	1 (7)	0 (0)	1 (7)	2 (19)	1 (8)	0 (0)	3 (27)
Total	4 (34)	5 (36)	0 (0)	9 (70)	7 (70)	4 (33)	2 (19)	13 (122)

Note: Number of weak students (about bottom 20% of a class) in parentheses.

TABLE 2

Summary Statistics of Evaluation Samples

	School B Nonmonetary			School A Monetary			School B Monetary		
	Control	Treatment	Diff.	Control	Treatment	Diff.	Control	Treatment	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	0.71	0.579	0.131	0.559	0.556	0.00327	0.7	0.606	0.0939
School A	0	0	0	1	1	0	0	0	0
Grade	4.35	4.05	0.302	3.53	4.08	–0.554	4.41	3.67	0.748*
Score2006	–1.08	–2.12	1.04**	–1.17	–2.13	0.964***	–1.36	–1.48	0.126
Score2007	–0.742	–1.17	0.423	–2.14	–2.21	0.0659	–0.393	–0.364	–0.0291
N	31	19		34	36		70	33	

Note: *Male* = a binary gender dummy (1 for male, 0 for female). *School A* = a binary school dummy (1 for school A, 0 for school B). *Grade* = a categorical dummy for school grade taking values in 2, 3, 4, 5, and 6. *Score2006* and *Score2007* = normalized scores for the baseline and evaluation scores, respectively.
* $P < 0.5$, ** $P < 0.01$, *** $P < 0.001$.

The evaluation sample for calculating the monetary program effect was defined similarly. It was composed of treated students and weak students in comparison classes from the same school. Because almost every grade in each school had at least one monetary program class, it was not necessary to restrict control students to those from the same grade/school combination. In total, the evaluation sample for monetary program effects in School A included 36 treated students and 34 control students (columns 4 and 5 of Table 2); the evaluation sample for monetary program effects in School B included 33 treated students and 70 control students (columns 7 and 8 of Table 2).

B. *Outcomes*

The main outcome of interest is whether the interventions improved the academic performance of the tutees. We used raw scores to measure academic performance. Data from the pretreatment final exam (fall 2006) and the posttreatment final exam (spring 2007) are available for both schools. The same grade in different schools uses different exams. The classes in the same grade of a particular school use the same exams. In what follows, all raw scores are normalized relative to the distribution of the pretreatment raw score in the comparison group in each school.[7]

Table 2 shows the descriptive statistics of the evaluation samples in three different cases: (1) School B, nonmonetary program vs. comparison group (or treatment vs. control); (2) School A, monetary program vs. comparison group; (3) School B, monetary program vs. comparison group. For example, the summary statistics (for gender, school, and grade dummies, and scores before and after the program *Score*2006 and *Score*2007) of the evaluation sample of the nonmonetary program effects is reported in columns 1 and 2 of Table 2, by treatment and control status. Column 3 reports the difference of columns 1 and 2, and the stars denote the level of statistical significance in the *t*-test.

Most of the pretreatment characteristics are balanced (i.e., with an insignificant *t*-test statistic) in Table 2. Our main confounding problem is that the pretreatment scores (*Score*2006) in the treatment groups seem to be systematically lower. In other words, the bottom students in the program classes are much weaker (in terms of the raw score) than their counterparts (in terms of the class ranking) in the comparison classes.

However, this does not mean that our evaluation design is fundamentally flawed. Randomization cannot guarantee complete pretreatment score balance, especially when the number of classes is relatively small, as in our case. We can further show that the unbalanced score is mainly a problem in the senior grades (results omitted). We always report the results of separate regressions for the junior and senior subsamples.

C. *Estimation Strategy*

Since we are interested in score improvement, a standard estimation strategy to use is differences-in-differences. If the score improvement trend is similar for the weak students in both the program and comparison groups, the coefficient of

[7] For each school, we subtracted the mean of the control group's pretreatment score and divided it by its standard deviation. The raw score is the sum of the scores for all subjects taught in a school. Chinese and math were taught in both schools. English was only taught in School A. Results based on separate subjects are similar and omitted.

the interaction term in the following differences-in-differences regression (i.e., α_3) would yield unbiased estimates of the treatment effects.[8]

$$Y_i = \alpha_1 Post + \alpha_2 Treatment + \alpha_3 DD + \beta Z_t + \varepsilon_i, \qquad (1)$$

where Y_i is score, *Post* is the time dummy that equals 1 for spring 2007 and 0 for fall 2006, *Treatment* equals 1 for the treated group and 0 for the corresponding control group, $DD = Post \times Treatment$ is the diff-in-diff variable, and Z_i contains other control variables including sex and class dummies.

V. MAIN PROGRAM EFFECTS FOR TUTEES

The tutees in at least some groups seem to have benefited from the programs. Figure 1 plots the difference in the posttreatment and pretreatment raw scores. The top, middle, and bottom rows correspond to three scenarios, respectively: (1) School B, nonmonetary program vs. comparison group; (2) School A, monetary program vs. comparison group; (3) School B, monetary program vs. comparison group. The program effects are quite significant in the top and middle panels. In other words, the tutees in the nonmonetary program and School A's monetary program experienced larger score improvement compared to those in the corresponding comparison groups. There seem to be no program effects in School B, suggesting the impact is context dependent. More rigorous and detailed regression estimation results are reported below.

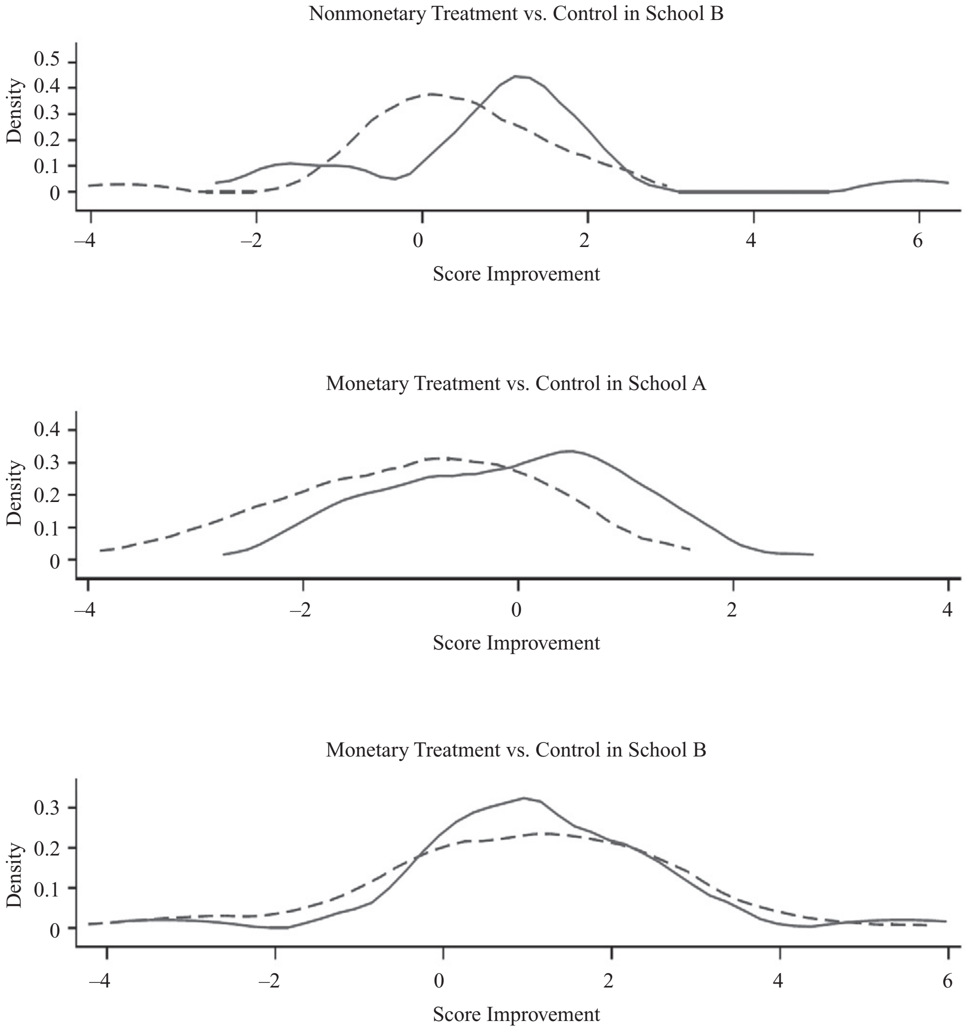
A. *Nonmonetary Program*

Even without monetary incentive, peer-tutoring seems to be effective. From Table 2, we can calculate from the summary statistics that the program effect is 0.612 in terms of the raw score improvement.[9] The *DD* estimation result (reported in column 1 of Table 3) is consistent with the back-of-envelope calculations. The first column reports the results using raw score as the outcome variable and controlling various covariates in equation (1). The *DD* estimator is 0.615 and statistically significant at the 5% level. This means that the nonmonetary program helped the tutees gain an additional score improvement of 0.615 standard deviations. Because of the relatively small sample size, we will not report results by different grades. Because of the panel data structure in diff-in-diff regression specification, the sample size reported in column 1 of Table 3 ($N = 100$) doubles that reported in Table 2 columns 1 and 2 ($N = 31 + 19 = 50$). The same is also true for the following diff-in-diff regression samples.

[8] If the weak students are less likely to make progress, this will strengthen our program results.

[9] $(-1.17 + 0.742)$-$(-2.12 + 1.08) = 0.612$

Fig. 1. Probability Distribution of Score Improvement



Nonmonetary Treatment vs. Control in School B

Monetary Treatment vs. Control in School A

Monetary Treatment vs. Control in School B

Note: The broken lines refer to the comparison groups. The solid lines refer to the program groups.

B. *Monetary Program*

We now move on to discuss the *DD* estimation results of monetary program effects (by school and seniority) reported in Table 3. Column 2 reports the results for School A. The *DD* estimator is 0.898, which is statistically significant at the 5%

TABLE 3

Monetary Program Effects on Tutees' Raw Scores

	School B Nonmonetary	School A Monetary			School B Monetary		
	All	All	Grades 2–3	Grades 4–6	All	Grades 2–3	Grades 4–6
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DD	0.615**	0.898**	0.785*	1.093*	0.155	0.874*	–0.131
	(0.260)	(0.369)	(0.360)	(0.500)	(0.412)	(0.411)	(0.707)
Treatment	–1.753***	–0.361	–0.313	–2.363***	1.039**	–0.536	–1.087*
	(0.270)	(0.440)	(0.358)	(0.414)	(0.458)	(0.361)	(0.605)
Post	0.335	–0.978***	–0.701*	–1.254***	0.963***	0.242	1.251***
	(0.184)	(0.217)	(0.358)	(0.212)	(0.300)	(0.382)	(0.360)
Male	–0.756**	–0.121	0.125	–0.279	–0.412*	–0.357	–0.451
	(0.320)	(0.164)	(0.174)	(0.246)	(0.209)	(0.358)	(0.269)
Class FE	YES	YES	YES	YES	YES	YES	YES
R-squared	0.343	0.511	0.491	0.485	0.392	0.284	0.381
N	100	140	58	82	206	76	130

Notes: 1. The regressions are specified in equation 1.
2. Outcome variable is raw score.
3. Standard errors (reported in the parenthesis) adjusted for intragroup correlation at class level.

* *P* < 0.1, ** *P* < 0.05, *** *P* < 0.01.

level. The 3rd and 4th columns report the results for grades 2–3 and 4–6 in School A, respectively. The estimated treatment effects are slightly weaker for the younger students in School A, though the difference is quite small. Both are statistically significant at the 10% level. The 5th column reports the results for School B. The estimated *DD* estimator is only 0.155, and is not statistically significant. This is largely because of the ineffectiveness of the program for the older students. The *DD* estimator for the younger or grade 2–3 students (reported in the 6th column) is 0.874, and is statistically significant at the 10% level. For the older or grade 4–5 students, the *DD* estimator (reported in the 7th column) is –0.131 and not statistically significant. In summary, except for the older students in School B, the monetary program is effective and helps the weak students gain an additional score improvement of about 0.9 standard deviations. The monetary program is about 0.5 times more effective than the nonmonetary program.

Why is the program effect weaker in School B, especially in its senior classes? Our interviews with the students suggest that there were significantly fewer actual peer-tutoring activities in School B's senior program classes. Teacher negligence and low incentives also appeared to be more serious in these classes as they failed to conduct the actual pairing of the assigned tutors and tutees in two of the program classes, forcing us to drop these two classes from the program class list. We are not sure about the cause of this noncompliance. It may be due to the differences in these two schools mentioned above. Kremer, Miguel, and Thornton (2009) also found that the compliance rate and the effects of their scholarship program was context dependent. Another possibility is that later-life-stage intervention is more difficult (Carneiro and Heckman 2003). The age difference is also consistent with findings in educational psychology (Cohen, Kulik, and Kulik 1982). We suggest that readers use caution in interpreting our results (including both monetary and nonmonetary program effects) based on School B, which is included in the current paper mainly as a reference.

There is no evidence that the program effects differ by gender or by the gender composition of the pairs. These results are not reported.

C. *Second Semester Results*

There is a large body of literature in economics and psychology that suggests that incentive contracts and material rewards may undermine voluntary cooperation.[10] We find no such evidence in our programs. This is not surprising, since there was a minimum amount of spontaneous peer-tutoring in the absence of our interventions.

The incentives provided to the tutors in our programs were removed in the fall semester of the 2007–8 school year in both schools. Testing long-run program effects after the removal of the financial incentive is a common practice to test for the presence of the motivational crowding-out effect (Kremer, Miguel, and Thornton 2009). Because the tutees knew their tutors were paid, the presence of tutor payment might have crowded out tutee motivation if they had a sufficiently high cooperative spirit. And the tutees may have become discouraged because of the withdrawal of the monetary program. It is also possible that payment may have reduced the tutors' intrinsic motivation to interact with the tutees. And they may have reduced the level of interaction with the tutees below the pretreatment level after the removal of the cash incentive.

We were only able to track the follow-up records in School A since the other school was hit by a sudden policy change around the middle of the spring semester when the local government asked many migrant children schools to close at the end of the academic year. Using the same pretreatment score, the program effect on raw scores is 0.80 in the second semester (significant at the 5% level), slightly smaller than the first-semester effect (see column 1 in Table 4). This suggests the existence

[10] For a discussion on economics see Mellström and Johannesson (2008). Educational psychologists have found that both extrinsic and intrinsic motivations could be important in peer-tutoring (O'Donnell and O'Kelly 1994; Topping 2005).

TABLE 4

Second-Semester Monetary Program Effects on Tutees' Raw Scores in School A

	All (1)	Grades 2–3 (2)	Grades 4–6 (3)
DD	0.800**	0.737*	0.893
	(0.286)	(0.359)	(0.499)
Treatment	–0.467	–0.876**	–1.716***
	(0.367)	(0.321)	(0.410)
Post	–0.348	–0.258	–0.454
	(0.283)	(0.325)	(0.501)
Male	–0.174	–0.220	–0.140
	(0.219)	(0.377)	(0.285)
R-squared	0.431	0.421	0.399
N	132	57	75

Notes: 1. The regressions are specified in equation 1.
2. Outcome variable is raw score.
3. Standard errors (reported in parentheses) adjusted for intragroup correlation at class level. Because of attrition, our sample size is slightly smaller than what is reported in Table 3.

* $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

of at least some long-run program gains for the tutees. The second-semester program impact appears to be more robust for the young students (column 2) than the older students (column 3), a pattern that is consistent with our main results. From Table 4, we can see the result is statistically significant for the young students at the 10% level, but for the older ones it is no longer statistically significant.

VI. CONCLUSION

This paper reports the results of a remedial education field experiment that used an incentive contract to encourage peer tutors to help their underperforming classmates in migrant children schools in China. Our experimental evidence suggests that paying peer tutors has a significant impact on tutees' grades in these schools. If the program studied here can be successfully scaled up across the country, the policy impact will be very significant because a huge number of migrant children lag behind students in the public education system.

Scaling up the program is also quite feasible. As both one-to-one peer-tutoring and merit-based scholarships are common practice in China, combining these two into an incentivized peer-tutoring program would be readily accepted by many teachers and schools. In our experience, teachers embraced our program with much enthusiasm. Therefore many migrant/rural schools may be willing to adopt such a program if there are such opportunities. At the same time, some funding for merit-based scholarship programs already exists. For example, Cyrus Tang Foundation has offered scholarships (largely based on merit) to 24,722 students in rural China from 2006 to 2010. If nongovernment organizations and government agencies can incorporate incentivized peer-tutoring into their existing merit-based scholarship and conditional cash transfer programs, smart but poor rural children can get some extra funding for their education, and underperforming rural students can improve their academic performance with the help of their peers.

Incentivized peer-tutoring is also free of the moral weakness believed to be inherent in merit-based scholarships. Many argue that the benefits flow disproportionately to well-off pupils and exacerbate inequality (Orfield 2002). This argument is particularly powerful for primary education, where equity seems to be a more important goal than efficiency. To get the prize in an incentivized peer-tutoring program, the tutors (who are also high-performing students) need to effectively strengthen their underperforming tutees, contributing to a unique equity-enhancing outcome.

REFERENCES

Banerjee, Abhijit V.; Shawn Cole; Esther Duflo; and Leigh Linden. 2007. “Remedying Education: Evidence from Two Randomized Experiments in India.” *Quarterly Journal of Economics* 122, no. 3: 1235–64.

Carneiro, Pedro, and James Heckman. 2003. “Human Capital Policy.” NBER Working Paper no. 9495. Cambridge, Mass.: National Bureau of Economic Research.

Cohen, Peter A.; James A. Kulik; and Chen-Lin C. Kulik. 1982. “Educational Outcomes of Tutoring: A Meta-Analysis of Findings.” *American Educational Research Journal* 19, no. 2: 237–48.

Dang, Hai-Anh, and Halsey Rogers. 2008. “The Growing Phenomenon of Private Tutoring: Does it Deepen Human Capital, Widen Inequalities, or Waste Resources?” *World Bank Research Observer* 23, no. 2: 161–200.

Glewwe, Paul, and Michael Kremer. 2006. “Schools, Teachers, and Education Outcomes in Developing Countries.” In *Handbook of the Economics of Education*, vol. 2, ed. Eric A. Hanushek and Finis Welch. Amsterdam: North-Holland.

Kremer, Michael; Edward Miguel; and Rebecca Thornton. 2009. “Incentives to Learn.” *Review of Economics and Statistics* 91, no. 3: 437–56.

Lai, Fang; Chengfang Liu; Renfu Luo; Linxiu Zhang; Xiaochen Ma; Yujie Bai; Brian Sharbono; and Scott Rozelle. 2012. “Private Migrant Schools or Rural/Urban Public Schools: Where Should China Educate Its Migrant Children?” Rural Education Action Project Working Paper 224. Stanford, Calif.: Stanford University.

Lazear, Edward P., and Sherwin Rosen. 1981. “Rank-Order Tournaments as Optimum Labor Contracts.” *Journal of Political Economy* 89, no. 5: 841–64.

Mellström, Carl, and Magnus Johannesson. 2008. “Crowding Out in Blood Donation: Was Titmuss Right?” *Journal of the European Economic Association* 6, no. 4: 845–63.

Mischo, Christoph, and Haag Ludwig. 2002. “Expansion and Effectiveness of Private Tutoring.” *European Journal of Psychology of Education* 17, no. 3: 263–73.

O'Donnell, Angela M., and James O'Kelly. 1994. "Learning from Peers: Beyond the Rhetoric of Positive Results." *Educational Psychology Review* 6, no. 4: 321–49.

Orfield, Gary. 2002. "Foreward." In *Who Should We Help? The Negative Social Consequences of Merit Aid Scholarships*, ed. Donald E. Heller and Patricia Marin. Cambridge, Mass.: Civil Rights Project, Harvard University: 25–40. http://www.eric.ed.gov/PDFS/ED468845.pdf (accessed July 22, 2012).

Robinson, Debbie R.; Janet Ward Schofield; and Katrina L. Steers-Wentzell. 2005. "Peer and Cross-Age Tutoring in Math: Outcomes and Their Design Implications." *Educational Psychology Review* 17, no. 4: 327–62.

Shastry, Gauri Kartini, and Leigh Linden. 2009. "Grain Inflation: Identifying Agent Discretion in Response to a Conditional School Nutrition Program." Mimeographed. http://ase.tufts.edu/econ/events/neudcDocs/SundaySession/Session35/GShastryGrainInflationIdentifyingAgent.pdf (accessed July 22, 2012).

Slavin, Robert. E. 2009. *Can Financial Incentives Enhance Educational Outcomes? Evidence from International Experiments*. Baltimore, Md.: Johns Hopkins University, Center for Data-Driven Reform in Education. http://www.bestevidence.org/word/finan_inc_Feb_4_2009.pdf (accessed July 22, 2012).

Topping, Keith J. 2000. *"Tutoring." Educational Practices Series No. 5. Brussels: International Academy of Education*. Geneva: International Bureau of Education.

———. 2005. "Trends in Peer Learning." *Educational Psychology* 25, no. 6: 631–45.

Wang, Zhaolong. 2006. "Yi Bang Yi Shi Zhuan Hua Wen Ti Xue Sheng De You Xiao Tu Jing" [One-to-one peer tutoring is an effective way to transform problem students]. *Du Yu Xie: Jiao Yu Jiao Xue Kan* 7: 70–76.

Xie, Qimei. 2008. "He Zuo Hu Lai Zai Ban Ji Jiao Xue Guan Li Zhong De Ying Yong" [The applications of peer-assisted learning in classroom teaching and administrations]. *Ke Xue Jiao Yu Jia* 8: 30–31.

Zhao, Shujie. 2009. "Yi Bang Yi Gong Cheng" [One-to-one peer tutoring projects]. *Xiao Xue Jiao Shi* 3: 45–54.