Research in Economics 72 (2018) 33–48

Contents lists available at ScienceDirect

Research in Economics

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

Heterogeneous effects of peer tutoring: Evidence from rural Chinese middle schools☆

Yang Song[a,*], George Loewenstein[b], Yaojiang Shi[c]

[a] Department of Economics, Colgate University, United States
[b] School of Social Decision Sciences, Carnegie Mellon University, USA
[c] Center for Experimental Economics in Education, Shaanxi Normal University, China

ARTICLE INFO

Article history:
Received 7 April 2017
Accepted 7 May 2017
Available online 12 May 2017

JEL classification:
I21
I24
I25

Keywords:
Peer tutoring
Group incentive
Heterogeneous effects
Mental health
Learning stress

ABSTRACT

Peer tutoring is a well-known type of peer-assisted learning, which has proven to be a cost-effective intervention. We designed a peer tutoring program that matches high-performing students as tutors to their low-performing classmates and provides non-monetary incentives for them to study together and improve the pair's academic performance. We implemented the program and tested the effects in rural Chinese middle schools. The program significantly improved the tutors' math scores and produced other benefits regarding study attitude and social behaviors. However, the program did not improve the tutees' math scores and instead augmented their learning stress. The most compelling explanation is that the set-up of the program brought to light the tutees' standing, by design, in the bottom half of their class.

1. Introduction

Peer-assisted learning has proven to be a cost-effective intervention (Levin et al., 1984), yielding gains in both academic and non-academic (social and communication) skills among students (Cohen et al., 1982; Rohrbeck et al., 2003; Topping, 2005). Classwide peer tutoring is one approach to peer-assisted learning, in which students in a class are paired to work together. Research indicates that this method can significantly improve students' performance in reading, spelling, and math at low cost (Fantuzzo et al., 1992; Fuchs et al., 1997; Greenwood et al., 1989). While most studies of peer tutoring have been conducted in developed countries, its cost-effectiveness is an even more desirable feature in resource-constrained developing countries.[1]

Building on a large body of literature on incentives to learn and peer effects, we designed and tested the impact of a peer tutoring program.[2] The program paired high-achieving students as tutors with their lower-achieving classmates of the

☆ We would like to thank Chu Yang and M. Najeeb Shafiq for their invaluable input at an early stage of the project. We are grateful to Lindsay Page and Saurabh Bhargava for their helpful comments. We would also like to thank Fei He and research assistants at CEEE, Shaanxi Normal University for their help with implementation of this study. This work was supported by the 111 Project (Grant No. B16031).

* Corresponding author
E-mail addresses: ysong@colgate.edu (Y. Song), gl20@andrew.cmu.edu (G. Loewenstein), shiyaojiang7@gmail.com (Y. Shi).

[1] See Kremer and Holla (2009) for a review on education interventions in developing countries.

[2] See Sacerdote (2011) for a review on peer effects in education and (Burke and Sass, 2013; Ding and Lehrer, 2007; Kang, 2007) for a few examples on causal classroom peer effects in the secondary school setting. Incentives to learn can take a variety of forms, including merit scholarship

http://dx.doi.org/10.1016/j.rie.2017.05.002
1090-9443/

same gender.[3] The program incorporated two types of incentives: snacks during peer tutoring time (input) and prizes to tutor/tutee pairs who increased their group ranking (output).[4] We implemented the program in two rural middle schools in Yulin, Shaanxi Province, during the fall semester of 2013, and examined its impact on student performance, attitudes towards school, and school retention rates. Baseline and post-intervention surveys on demographics, family background, mental health were collected, and standardized tests were conducted.

The peer tutoring program had both intended and unintended effects. On the positive side, tutors experienced an improvement of 0.41 standard deviations in standardized math scores. In addition, none of the tutors in the treatment group dropped out of school during the semester, compared with three out of the 99 students in the control group (who would have been tutors had they been in the treatment group). The program also led to improvements among tutors in mental health measures, attitudes toward schooling, and incidences of misbehavior (arguments and fights with classmates and absences).

However, contrary to the program's goal, tutees exhibited no gains in math scores and a decline of 0.27 standard deviations in their self-reported mental health scores after participating in the program. Breaking the mental health scores into subcategories, we found that the worsened mental health scores were mainly due to higher learning stress. The most compelling explanation is that the set-up of the program brought to light the tutees' standing, by design, in the bottom half of their class.

To our knowledge, this study is the first to examine how structured peer tutoring affects students' mental health. The increase in tutees' learning stress could be due to the publicity of their standing in the bottom half of their class brought to light by the program. In the post-intervention survey, 32.2% of tutees reported that they felt a little or very embarrassed by the program, as opposed to 16.7 percent of tutors.

The results underscore the importance of how peer tutoring programs are implemented. Although some previous studies examined the social and emotional effects of peer tutoring, these earlier studies mainly focused on the program's impact on attitudes towards school and classmates and social behavior (Scruggs and Mastropieri, 1998; Sutherland et al., 2000; Tolmie et al., 2010).[5] The positive effects on tutors are consistent with previous findings by Webb and Farivar (1994) that giving help and providing answers to peers is associated with improved mathematics achievement. Maher et al. (1998) also found that, when disruptive students served as peer tutors, their achievement and behavior improved.

The closest program design to our study is Li et al. (2014), who conducted a randomized trial with primary school students in China and offered pairs of high and low achieving students group incentives for learning. There are two main differences between their intervention and ours. First, we incentivized for pair improvement, while they incentivized for tutees' improvement. Second, in their study, the top quarter of students were matched to sit with and help the bottom quarter of students to improve, and therefore the middle 50% students were not involved; in our study, all students were part of the peer tutoring program. In contrast to our findings, Li et al. (2014) did not find significant improvement among the helpers, but tutees improved their test scores by approximately 0.265 standard deviations. The positive effects on tutees and insignificant effects on tutors in their study may partly reflect the fact that they incentivized tutees' improvement rather than (as we did) the improvement of both tutor and tutee.

In the remainder of this paper, Section 2 describes the design of the peer tutoring program and presents some background information on the schools where the study was conducted. Section 3 evaluates the effects of peer tutoring program quantitatively and presents some qualitative feedback from post-intervention surveys. Section 4 discusses some possible explanations for our results and proposes program modifications that could potentially prevent the unintended effects on tutees.

2. Program design and implementation

In rural Northwestern China, more than 14% of middle school students drop out during their Nine Year Compulsory Education (Yi et al., 2012). In light of the educational and social/emotional issues faced by students, many of whom live away from their families while attending school, peer tutoring may address the diverse needs of tutees, while also benefiting tutors, who could derive pleasure and pride from helping their peers and learn from teaching.

2.1. Program design

The peer tutoring program matched top students as tutors to lower-performing students and provided incentives for the two to study together, improve their academic performance, and complete their education.

(Angrist et al., 2009; Cornwell et al., 2006; Kremer et al., 2009), rewards for performance improvement or number of books read (Fryer, 2011), and conditional cash transfers (Rawlings and Rubio, 2005). Gneezy et al. (2011) provide a review of these studies; incentives for attendance and enrollment are generally effective, while incentives for achievements bring mixed results.

[3] Rohrbeck et al. (2003) found that same-gender peer-assisted learning studies have a greater effect size on average than mixed-gender ones.

[4] Previous research has documented benefits from introducing team-based incentives for education, especially for disadvantaged groups (Blimpo, 2014; Li et al., 2014).

[5] Sutherland et al. (2000) reviewed eight experimental studies of the effectiveness of cooperative learning for students with emotional and behavioral disorders, and found mixed results.

To create tutor/tutee pairs, students in each class were ranked within gender groups according to their overall academic performance from the previous semester (based on their performance in reading, math, English, politics, history, geology, and biology), and divided into two equal-sized groups of those in the top and bottom 50th percentiles. We then matched same-sex pairs of students within their percentile group. For example, the top female student was matched with the female student who was just below the 50th percentile within the female students. If there were 40 students in a class (20 boys and 20 girls), the girl ranked first would tutor the 11th ranked girl, and the 10th ranked girl would tutor the 20th ranked girl.[6] The ranking gap between the tutors and tutees is, by design, the same for every pair. We arranged the pairing in this way to maximize the difference and consistency of the difference in academic performance between the tutor and tutee.

The program incorporated two types of incentives to encourage students to spend time together (input) and to improve the pair's class rank (output). To encourage the pairs to study together, we provided tasty, healthy snacks (like rice cakes, dairy candies, or fiber bars) for class teachers to reward both tutors and tutees when they studied together. Groups were required to study at least 30 minutes per day to get a daily snack, and at least four times a week to get a weekly bonus snack. In addition, rewards, such as notebooks and pens, were available to pairs based on the improvement in the average ranking of tutor and the tutee in the school's monthly exams that cover all subjects.

We chose to incentivize the students to improve the ranking improvement in the schools' own comprehensive exams instead of that our own post-intervention standardized math scores. This design was intended to encourage students to closely follow and grasp what they were taught in school, instead of narrowly focusing on math at the expense of other subjects to get the prize.

2.2. Implementation

We tested the program in two rural middle schools in Yulin, Shaanxi Province, during the fall semester of 2013. Students in these schools are from low-income rural families. According to our baseline survey, 59.6% of the fathers and 75% of the mothers have six years of education or less, and 91.5% of fathers and 94.5% of mothers have nine years of education or less. 40% of students come from households that lack running water.

For the peer tutoring program, we targeted seventh grade students enrolled in No. 5 Middle School and eighth grade students in Liangzhen Middle School. Eighth grade students at No.5 Middle School and seventh grade students at Liangzhen Middle School comprised the control groups. In total, there were six classes in each group, with 242 students in the treatment group and 216 students in the control group.

The treatment assignment reflected the sample constraints and other considerations. First, although middle schools in China include seventh, eighth, and ninth grades, we did not include the ninth grade in the study, because the schools did not want our research team to interfere with the ninth grade students, who were busy preparing for the high school entrance exam. Second, we chose to treat one grade in each school, instead of using one school for treatment and the other for control, in part because one school is almost double the size of the other; if we used one as the treatment school, our treatment and control groups would be unbalanced in size. A second consideration is that, even though two schools may be similar to each other, there still might be school-specific factors that would influence the outcomes and thus prevent us from isolating the effects of treatments.

We conducted baseline surveys and standardized math tests in late September 2013, before the program was implemented. We administered a standardized math test to evaluate the impacts on academic performance, since the final exam test scores from different schools are not comparable. The math exams were designed by local educators for different grades to evaluate their understanding of the common curriculum. Post-treatment tests were slightly more difficult than the baseline tests. Since neither compensation for teachers nor prizes for students were dependent on performance in the math tests, no one had the incentive to cheat or make extra efforts to prepare for these tests.

After collecting data from the surveys and math tests, the researchers explained the peer tutoring program to teachers and students, including the peer matching protocol. After the initial visit, the researchers visited the schools twice during the fall semester to replenish the supply of snack rewards and prizes for improvement. We conducted an evaluation survey and standardized math test in late February 2014. The expenses for snacks, prizes for improvement, and compensation for extra work put in by the teacher to check the notes and give snack prizes equaled approximately $2,100, $150, and $100 USD, respectively, totaling to approximately $2,350 USD.

2.3. Balance check: pre-treatment characteristics

The baseline information we collected included three main parts: 1. A student survey with questions on family backgrounds, attitudes towards school, self-efficacy and emotion management, predictions about who was most likely to drop out in class, etc. A mental health survey was included in the baseline survey. We computed the mental health score based on answers to ninety questions on learning stress, fear, impulsiveness, etc.; 2. A standardized math test with twenty-five

[6] In the cases of odd numbers of female (male) students, the top-ranked student forms a three-people group with the middle- and the bottom-ranked student. For example, if there were 21 girls in one class, the 1st ranked girl would tutor both the 11th and 21st ranked girls. We checked whether group size matters for treatment effects and did not find statistically significant results.

Table 1
Baseline student characteristics and balance check.

	Tutee			Tutor		
	Control	Treatment	Diff	Control	Treatment	Diff
Female	0.47	0.52	0.0502	0.475	0.5	0.0253
	(0.501)	(0.502)	(0.0648)	(0.502)	(0.502)	(0.0684)
Class size	36.47	40.468	3.998***	36.556	40.559	4.004***
	(4.448)	(2.676)	(0.476)	(4.529)	(2.652)	(0.516)
Grade 7 Std Math	−0.689	−0.476	0.213*	0.633	0.76	0.127
	(0.452)	(0.748)	(0.124)	(0.765)	(1.119)	(0.198)
Grade 8 Std Math	−0.864	−0.224	0.640***	0.488	0.504	0.0156
	(0.414)	(0.804)	(0.109)	(0.773)	(1.104)	(0.178)
Mental Health	61.63	59.25	−2.375	62.25	61.84	−0.409
	(14.176)	(14.129)	(1.888)	(14.667)	(12.808)	(1.946)
Poverty Program	0.353	0.315	−0.0389	0.404	0.333	−0.0707
	(0.48)	(0.466)	(0.105)	(0.493)	(0.473)	(0.0998)
Father Education	2.094	2	−0.094	2.202	2.246	0.0437
	(1.008)	(1.004)	(0.154)	(1.078)	(1.054)	(0.188)
Mother Education	1.701	1.782	0.0814	1.808	1.822	0.140
	(0.94)	(0.984)	(0.182)	(0.986)	(0.957)	(0.141)
Absence	1.077	1.149	0.0723	1.747	1.822	0.102
	(0.683)	(0.723)	(0.130)	(0.747)	(0.833)	(0.162)
Obs	117	124		99	118	

Note: Students with below-mean baseline math scores in the control classes consist of the control group for tutees in the treated classes. Similarly for tutors. The "Control" and "Treatment" columns show the means and standard deviations in parentheses. The "Diff" column shows the difference and robust standard errors in parentheses, with significance level indicated by ***p< 0.01, **p< 0.05, and *p< 0.1. Baseline math scores are standardized within grade. Mental health score is calculated based on 100 questions, with higher score meaning better mental health condition. Parental education is categorical as follows: less than elementary (1), elementary (2), middle school (3), high school (4), vocational college (5), college graduates or above (6). Absence is according to self-reported frequency per semester.

multiple-choice questions; and 3. A teacher survey asking about their qualification, effort, values, and predictions about which students were most likely to drop out.

Table 1 presents some baseline characteristics of tutors and tutees in the treatment and control groups. We use students who were ranked in the upper- and lower-half in the control group as the counter-factual groups for tutors and tutees. More specifically, we assign a placebo tutor or tutee dummy for each student in the control group: those who scored above median among the students of the same gender in class are assigned as placebo tutors; those who scored below are placebo tutees.

Most of the pre-treatment characteristics are balanced between two groups, including mental health score, parental education, family size, poverty program participation, and classroom behavior. However, tutees in the treatment group have higher baseline standardized math scores, which could bias the treatment effects upward or downward. Compared with lower-achieving students in the control group, it could be more difficult for the treatment group tutees to make more progress, because it may be easier to improve from 40 to 60 than to improve from 60 to 80. On the other hand, for tutees with a better foundation of knowledge, their tutors may have an easier time to help them catch up. Therefore, the direction of bias, if any, is unclear.

The other major unbalanced variable is class size, which is on average 4 students larger (40.5 versus 36.5) for treatment group students. Since previous studies have shown that a smaller class size generally improves learning outcomes (Krueger, 1999), a larger class size in the treatment group would probably bias against a positive treatment effect.

Other than the aforementioned baseline comparisons, we performed extra balance checks and present the results in Table A1. On a scale from 1 to 10 in agreement with the statement "I like going to school", tutees reported a 0.7 lower score relative to their counterparts in the control group, which may also create a bias against the positive treatment effect on tutees' academic and mental health improvement. Tutees reported fewer arguments with classmates, and tutors reported less fighting with classmates, compared to the control group. There were no statistically significant differences between the treatment and control groups for the following teacher characteristics: years of teaching experience, public qualification, years of obtaining the qualification, years being the class teacher, or any other measured attribute (see Appendix for details).

We conducted an omnibus test for all covariates (Hansen and Bowers, 2008) and rejected that the classes in the treatment group and those in the control group are the same. As we discussed above, some unbalanced factors (such as class size and study attitude) could bias the treatment effects downward, while some other factors (such as math score and misbehavior) could lead to a higher or lower effect. Since we do not have perfectly balanced baseline characteristics between treatment and control groups, especially for tutees, we include all controls for baseline characteristics in our regressions.

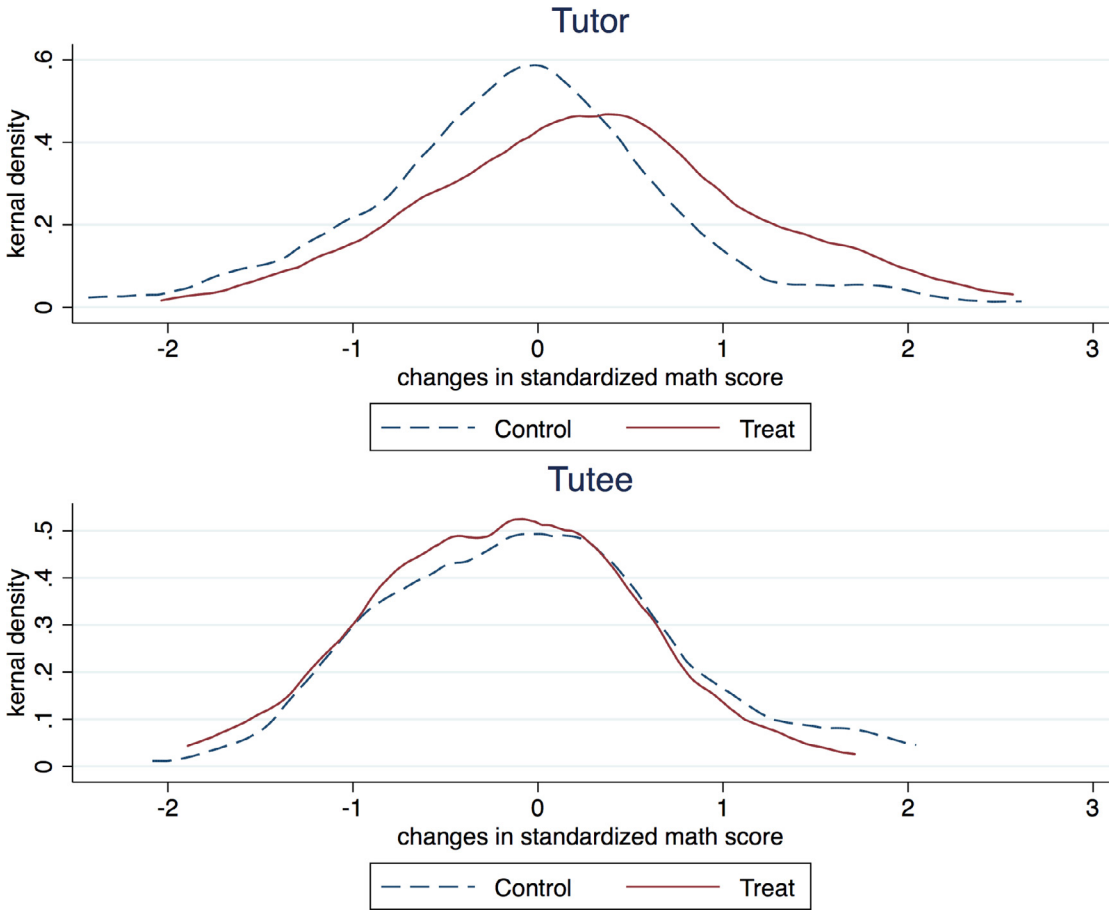

Fig. 1. Distributions of Change in Standardized Math Scores. Notes: Each figure plots residuals from regressing changes in standardized math scores on baseline characteristics. The top figure only uses the tutor/placebo tutor sample, and the bottom one only uses tutee/placebo tutee sample. Math scores are standardized at the grade level with a mean of 0 and a standard deviation of 1 for pre- and post-intervention separately. Placebo tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. The p-values of Kolmogorov–Smirnov equality tests for tutor and tutee graphs are 0.002 and 0.711, respectively.

3. Evaluating the peer tutoring program

We estimate the treatment effects on a variety of outcomes, including standardized math scores, mental health scores, and other non-academic outcomes such as attitudes and misbehavior. We also summarize the findings from subjective evaluations and suggestions in post-intervention surveys.

First, to obtain a graphical understanding of our key outcome variables, we plot the kernel density graphs of changes in standardized math scores and mental health scores for tutors and tutees. Since the baseline math scores were not balanced between the control and treatment group, we regress *changes in math and mental health scores* on the baseline characteristics of students and plot the residuals by treatment and control group. Fig. 1 shows that tutors in the treatment group significantly improved their standardized math scores, compared with the placebo tutors in the control group. The Kolmogorov-Smirnov test rejects the equality of changes in standardized math scores between treated tutors and control group tutors, with a p-value of 0.002. As for tutees, however, the two distributions are not statistically significantly different from one another (p-value 0.711).[7]

Since mental health scores are balanced in the baseline, we plotted raw data of first differences in mental health scores (post-pre) in Fig. 2.[8] For tutors, the distribution of control group concentrates at around zero, suggesting that the placebo tutors demonstrated little change in mental health score. Tutors, on the other hand, exhibit more changes, but on average still center around zero. Some tutors experienced improvement in their self-reported mental health scores, while others show some decline. Changes in mental health scores for tutees have larger range and more variation. Distribution for placebo

[7] Raw data without first differences is plotted by tutor/tutee and treat/control in Fig. A1.

[8] We also made similar residual plots for mental health scores as for standardized math scores in Fig. A3, which look similar to Fig. 2.

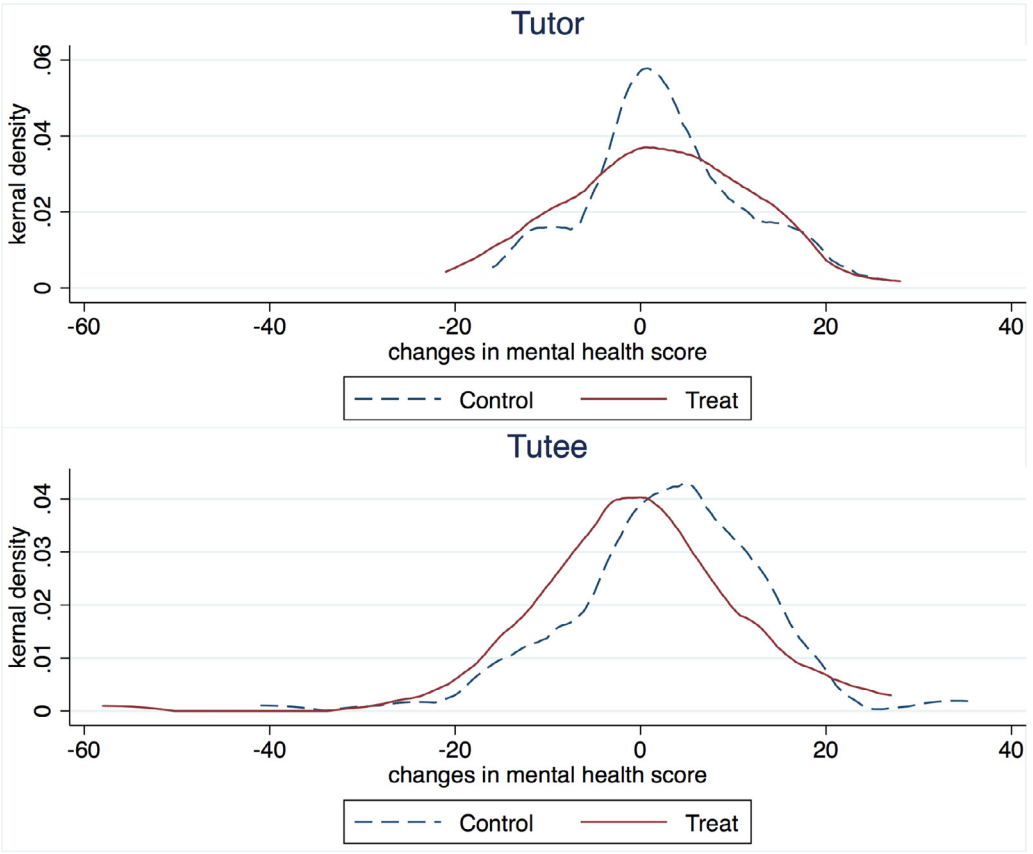

Fig. 2. Distributions of Change in Mental Health Scores. Notes: Each figure plots changes in mental health scores of the treatment group and that of the control group. Placebo tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. Mental health scores have an average of around 62 and a standard deviation of around 14.6, with higher scores indicating better mental health.

tutees centers to the right of zero, indicating a positive change in their mental health score. Meanwhile, the distribution for tutees centers around zero with a long negative tail, and is to the left of the distribution for placebo tutees.[9]

3.1. Main results

The main regression equation we use in this paper is:

$$Y_{ic} = \beta_0 + \beta_1 T_c + \beta_2 \mathbf{X_{ic}} + \mu_{ic} \quad (1)$$

where Y_{ic} is a set of outcome variables, including post-intervention standardized math scores and mental health scores. T_c is a dummy variable for program treatment status, which equals one if class c is treated. β_1 is the coefficient of interest. X_{ic} represents a set of 27 pre-program characteristics, including standardized math score, demographics, family background, baseline study attitude, misbehavior, mental health survey scores, etc. μ_{ic} is independent across different clusters c but correlated within clusters (classes).[10]

Table 2 presents estimates for β_1, treatment effects for standardized math scores and mental health survey scores. Each column represents a separate regression. The first two columns look at standardized math scores as the outcome variable and find different effects for tutors and tutees. While tutors experienced a 0.411 standard deviations improvement in math scores, the treatment effect for tutees' math scores is statistically insignificant.

As for mental health scores, tutees reported a significantly worse mental health survey score (0.27 standard deviations), while tutors stayed the same. This overall self-reported mental health score can be decomposed into subcategories to enable us to better understand the sources of deteriorated mental health scores of tutees.

Panel A in Table 3 presents the treatment effects on subcategories of the mental health survey score on tutees and tutors. We see that the worsened total mental health survey scores of tutees are mainly driven by higher learning stress. More daily

[9] Raw data without first difference is plotted in Fig. A2.

[10] Because of the small number of clusters we have, we also used the wild bootstrapping method (Cameron et al., 2008) and verified that our results would still hold.

Table 2
Treatment effects on math and mental health.

Dependent Var	Standardized Math		Mental Health Score	
	(1)	(2)	(3)	(4)
Subsample	Tutor	Tutee	Tutor	Tutee
mean	0.46	−0.43	64.32	61.45
s.d.	0.97	0.83	13.85	15.22
Treatment Effect	0.411*	−0.409	−0.0952	−4.115***
	(0.207)	(1.253)	(0.0858)	(1.279)
Observations	208	198	223	211
R-squared	0.330	0.660	0.310	0.567

Robust standard errors are clustered at the class level in parentheses. Significance level ***p< 0.01, **p< 0.05, *p< 0.1. All regressions include a set of baseline controls, including standardized math score, mental health survey outcomes, demographics, family characteristics, and school attitudes and activities. Standardized math score is constructed to have a mean of 0 and a standard deviation of 1 for each grade in each exam. Mental health score is calculated according to a 100 question survey, with higher score being better mental health condition. Results on the full specification are reported in the Appendix.

Table 3
Treatment effect on other variables of possible interests.

Panel A: Treatment Effect on Itemized MHT score								
	(1) Learning stress	(2) Social stress	(3) Anti-social	(4) Self-guilty	(5) Over-sensitive	(6) Physical-signs	(7) fear	(8) impulsiveness
Tutor	0.150	−0.497**	−0.545**	0.952**	0.0410	−0.150	0.172	−0.219
	(0.280)	(0.209)	(0.179)	(0.397)	(0.151)	(0.192)	(0.313)	(0.215)
Tutee	1.027**	0.152	0.216	0.577	0.194	0.365	0.350	0.418
	(0.413)	(0.290)	(0.243)	(0.382)	(0.279)	(0.326)	(0.239)	(0.255)
Panel B: Treatment Effect on Study Attitude								
	(1) like school (1–10)	(2) class participation	(3) ask question	(4) read	(5) extra exercise	(6) class leader		
Tutor	0.365**	−0.892***	0.111	0.106	0.0886	0.0935**		
	(0.158)	(0.282)	(0.0798)	(0.0723)	(0.0669)	(0.0400)		
Tutee	−0.336	−0.0151	−0.0882	−0.0836	0.124*	0.0197		
	(0.409)	(0.581)	(0.0908)	(0.0917)	(0.0616)	(0.109)		
Panel C: Treatment Effect on Behavior								
	(1) bullied	(2) argument	(3) fight	(4) late hw	(5) late class	(6) absence		
Tutor	−0.0735*	−1.433*	−0.440**	−0.0465	−0.0909	−0.555***		
	(0.0342)	(0.755)	(0.154)	(0.159)	(0.255)	(0.117)		
Tutee	0.107	0.363	−0.153	−0.317	0.00984	−0.121		
	(0.0904)	(0.258)	(0.225)	(0.355)	(0.271)	(0.116)		

Robust standard errors are clustered at the class level in parentheses. Significance level ***p< 0.01, **p< 0.05, *p< 0.1. All regressions include a set of baseline controls. Reported are the coefficients for treatment dummies. In Panel A, a positive coefficient means a worse problem in that category. For example, learning stress is higher in the treatment group than in the control group for tutees. In Panels B and C, "class participation" asks for daily frequency; "ask question", "read", "extra exercise", and "bullied" are categorical variables from never (0), to sometimes (1), to often (2); other than "bullied", all other variables in Panel C indicate frequency per semester.

study time and the stress from wanting to improve performance and ranking in return for their tutors' help may have led to the higher learning stress of tutees.

Tutors experienced less social stress and were less anti-social after the peer tutoring program. The interactions within pairs may have helped tutors to relieve some social pressure and improve their social skills. However, tutors showed higher levels of guilt, which may have resulted from their inability to help resolve the tutees' problems. The positive and negative effects of these three subcategories cancel out, which gives us an insignificant change in overall mental health scores for tutors.

The comprehensive survey also allows us to look at a larger set of non-academic outcome variables, including study attitude, effort, and misbehavior. Results are reported in Panel B and C in Table 3. Most of these estimates are insignificant for tutees, other than completing significantly more exercise problems. For tutors, the peer tutoring program increased their liking for school ("On a scale of 1 to 10, how much do you like going to school?"), decreased the frequency of being bullied or getting into arguments or fights, and lowered absence rates. This may reflect that tutors became more respected within the classes because of their role in the peer tutoring program.

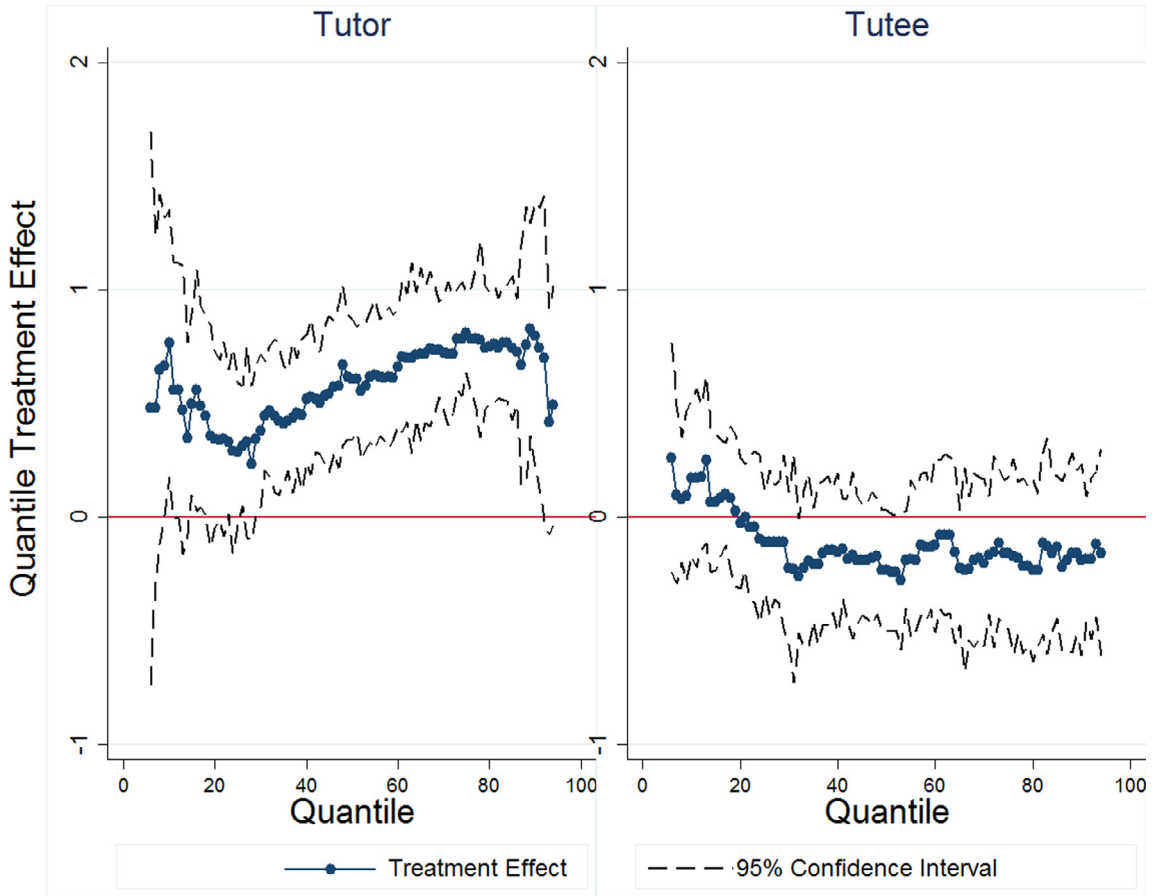

Fig. 3. Quantile Treatment Effects on Standardized Math Scores Notes: Math scores are standardized at the grade level with a mean of 0 and a standard deviation of 1. Tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. The solid line shows the quantile treatment effects estimates, and the dashed lines are the 95% confidence intervals.

Lastly, to understand what might be driving the observed effects of the program, we looked separately at students in the treatment group and tested if the following factors were correlated with more positive treatment effect on math score: different levels of engagement and satisfaction in the program (choosing the same partner again, feeling hopeful, feeling anxious, deeming the program helpful, etc.), peer baseline academic performance, difference between self and peer baseline academic performance, whether the class teacher teaches math, teacher's years of experience, etc. However, after controlling for baseline score and a few characteristics, none of these factors were statistically significantly correlated with post standardized math scores. Note, however, that the insignificant results could also be due to the small sample size.

3.2. Heterogeneous treatment effects

To obtain an understanding of how distributions of math and mental health scores changed, we ran quantile regressions to quantify the treatment effects at different percentiles of student academic performance and mental health scores.

$$Y_{ic} = \alpha_0^p + \alpha_1^p T_c + \alpha_2^p X_{ic} + \eta_{ic}^p \tag{2}$$

where Y_{ic} is a set of outcome variables including standardized math scores and mental health survey scores, T_c is the treatment dummy, X_{ic} is a list of baseline control variables, and $0 < p < 1$ indicates the proportion of the population having scores below the pth percentile.

Fig. 3 presents the results on the quantile treatment effects on math scores, separately for tutees and tutors. Each quantile regression estimate is plotted as a dot on the solid line in the middle, which shows the treatment effect on that specific quantile of the distribution. Dashed lines show the 95% confidence intervals. The peer tutoring program improved standardized math scores for almost all tutors. Although we see no statistically significant benefits on tutees' math performance, tutees at the lowest percentiles had slightly higher point estimates of their math scores.

Tutors generally gained more from the program when they had better academic performance, although the top 10% of students may not have benefited as much. More specifically, the gains in post standardized math score are statistically

Fig. 4. Quantile Treatment Effects on Mental Health Scores Notes: Mental health scores have an average of around 62 and a standard deviation of around 14.6. Tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. The solid line shows the quantile treatment effects estimates, and the dashed lines are the 95% confidence intervals.

insignificant below 30 percentile, then they increased for tutors from 0.3 standard deviations at the 30th percentile to 0.8 standard deviations at the 80th percentile, but they decreased to 0.5 standard deviations for those at the 90th percentile among tutors.

A plausible explanation for the heterogeneous effects on tutors' gains in math scores is that tutors experienced different levels of complementarity between the time spent on peer tutoring and time spent on self-improvement. Tutors at lower quantiles may have struggled to understand the materials clearly themselves, while tutors at higher quantiles were able to help tutees better and benefited more from clarifying questions. For the top 10% of the tutors, they may have already grasped the easier part of the materials thoroughly, and they benefited less from clarifying those to tutees.

The quantile treatment effects on mental health scores are shown in Fig. 4, where we see no significant effect on tutors but negative effects on tutees, mainly for the students at around the 75th percentile of the mental health score distribution. This suggests that worsened mental health scores for tutees mainly came from those with better mental health conditions. The magnitude of worsened mental health score is around 0.28 standard deviations.

3.3. Treatment effects by gender

Understanding how the program impacts different subgroups is important for assessing whether it is likely to be more successful with female or male students. We add to the baseline regression Eq. (1) interaction term $T_c{}^*X_{ic}$ of the treatment dummy variable with gender.

$$Y_{ic} = \beta_0 + \beta_1 T_c + \beta_2 X_{ic} + \beta_2 T_c * X_{ic} + \epsilon_{ic} \tag{3}$$

Previous research has found distinct responses by female and male students to educational interventions (Angrist et al., 2009 for example). As reported in Table A3, we do not find any significant gender difference in treatment effect on math or mental health. However, the positive signs on the interaction terms on treatment and female for standardized math scores, for both tutors and tutees, is consistent with previous literature. The insignificance of results may be due to the small sample size.

Table 4
Subjective survey evaluation.

Part A: Is the peer mentoring helpful?			
		Tutor % student	Tutee % student
helpful			
	Not at all	14.91%	3.48%
	A little	57.02%	41.74%
	Very much	28.07%	54.78%
Part B: Does peer tutoring makes you feel ... ?			
		Tutor % student	Tutee % student
proud			
	Not at all	51.75%	39.13%
	A little	39.47%	52.17%
	Very much	8.77%	8.70%
anxious			
	Not at all	55.26%	52.17%
	A little	42.11%	43.48%
	Very much	2.63%	4.35%
embarassed			
	Not at all	83.33%	67.83%
	A little	14.91%	31.30%
	Very much	1.75%	0.87%
hopeful			
	Not at all	17.54%	10.43%
	A little	55.26%	59.13%
	Very much	27.19%	30.43%
frustrated			
	Not at all	78.07%	74.78%
	A little	19.30%	22.61%
	Very much	2.63%	2.61%

3.4. Follow-up measures

In addition to the demographic, behavioral, and mental health questions in the baseline survey, we also included peer program evaluation questions in the post survey. More than 90% of students found the program to be at least a little helpful (see Table 4). Around 70% of students reported that they would choose the same partner if they were to complete the program again.

When we separately look at tutors and tutees, an interesting pattern emerges. Despite the largely positive results for tutors and insignificant/negative results for tutees in the quantitative evaluation, more tutees than tutors reported that the program is "very helpful" (55% versus 28%) and that they would choose the current partner again (84% versus 54%).

We also asked five questions about participants' feelings towards the peer tutoring program, including pride, anxiety, embarrassment, hope, and frustration. Mostly the answers were positive. Few students reported that they felt "very" anxious, embarrassed, and frustrated, and most of those who did were lower-performing students. Thirty-two percent of tutees reported that they felt a little or very embarrassed by the program, as opposed to sixteen percent of tutors.

In the evaluation survey, students also reported on tutoring timing, frequency, length, and the main subject discussed. Most students (over ninety percent) reported having participated in peer tutoring more than four times a week, the minimum requirement to get the weekly bonus snack. Most peer tutoring activities happened between classes or during morning and night self-study sessions, and usually lasted for twenty to thirty minutes per day. We tested whether the frequency and duration of tutoring can help explain the effects on academic performance or mental health score of the tutor/tutee, but did not find any statistically significant evidence.

4. Conclusions

In what may be the first evaluation of peer-assisted learning in a developing country setting, we designed and tested a peer tutoring program that paired high-performing and low-performing students and incentivized them to study together and improve as a group. The most striking finding is that, contrary to our expectations and intentions, tutees did not experience improvement in academic performance, but instead, suffered from lower self-reported mental health scores. On the positive side, tutors benefited on multiple dimensions, including higher standardized math score, lower social stress, lower dropout rate, more favorable attitudes towards school, less absence, and less misbehavior.

Although our intervention incorporated incentives for both study effort and academic improvement, the program, as implemented, turned out to place far more emphasis on snack prizes for peer tutoring time than on prizes for group improve-

ment. This was conveyed to us during the interviews with teachers and in student comments on post-treatment surveys. Other than the theory proposed by Fryer (2011), that students may not know the education production function to respond to outcome incentive effectively, it is also possible that students were present-minded and hence more motivated by the immediate and frequent availability of snacks. Students may not have cared enough about the quality of peer tutoring to get the rewards for group performance improvement as much as they cared about putting in the minimum amount of tutor/tutee interaction to get the snacks.

Another possible explanation for the lack of improvement in tutees' performance is that the rewards for improving the tutees' performance may not have been sufficient to motivate students to focus on tutees' learning progress. This may be an important reason why our results are different from those of previous studies, which found that using monetary prizes as group incentives led to improved test scores. As mentioned in the introduction, the comparison with Li et al. (2014) suggests that more emphasis on tutees' improvement could possibly have positive effects on their academic performance.

Since few of the previous studies examined the impacts of peer tutoring programs on student mental health, our findings call for caution when implementing programs which paired students based on their performance. Publicity of being in the lower half of the class could introduce problems of stigma and bring higher learning stress for tutees. How a program is framed and delivered to students directly affects their perception of themselves and others, and can affect the program's influences on student outcomes. A cross-age peer tutoring program, which pairs students in higher grades with students in lower grades, might avoid the stigma effect for tutees while still providing students in the higher grades, with the benefits of tutoring. Yet, in such a program, it is unclear if we would still observe the positive effects on tutors, since tutoring materials would be less synced with tutors' own learning, and therefore may take time away from studying for themselves.

Appendix. More Details on Group Incentives

The peer tutoring program incentivized students to spend time together (input) and improve the pair's class rank (output). To encourage the pairs to study together, we provided tasty, healthy snacks (like rice cake, dairy candies, or fiber bars) for class teachers to reward both tutors and tutees if they studied together. They could study between classes, at morning

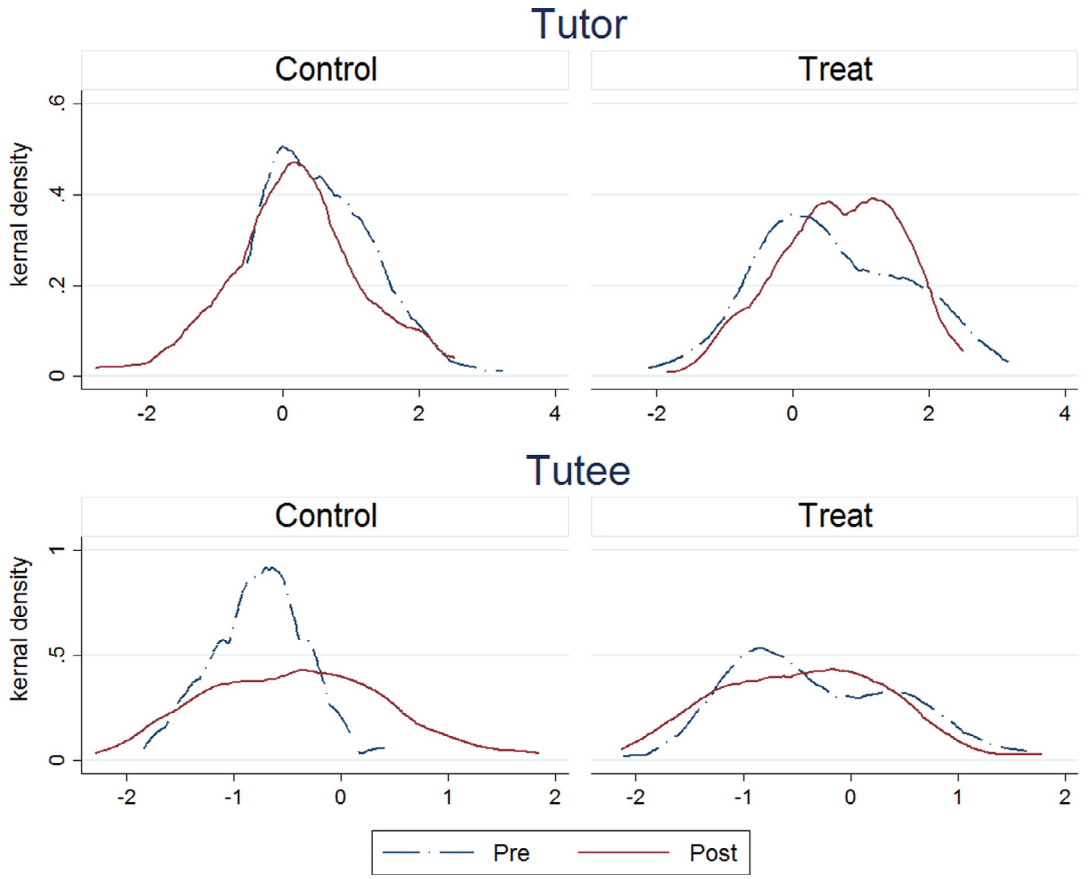

Fig. A1. Distributions of Standardized Math Scores Notes: Four plots show kernel density distributions of standardized math scores for four different groups of students, divided by their tutor/tutee status and treatment/control assignment. Placebo tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. In the baseline, tutees had significantly higher math scores than the placebo tutees. After the program, they still outperformed placebo tutees but the gap decreased.

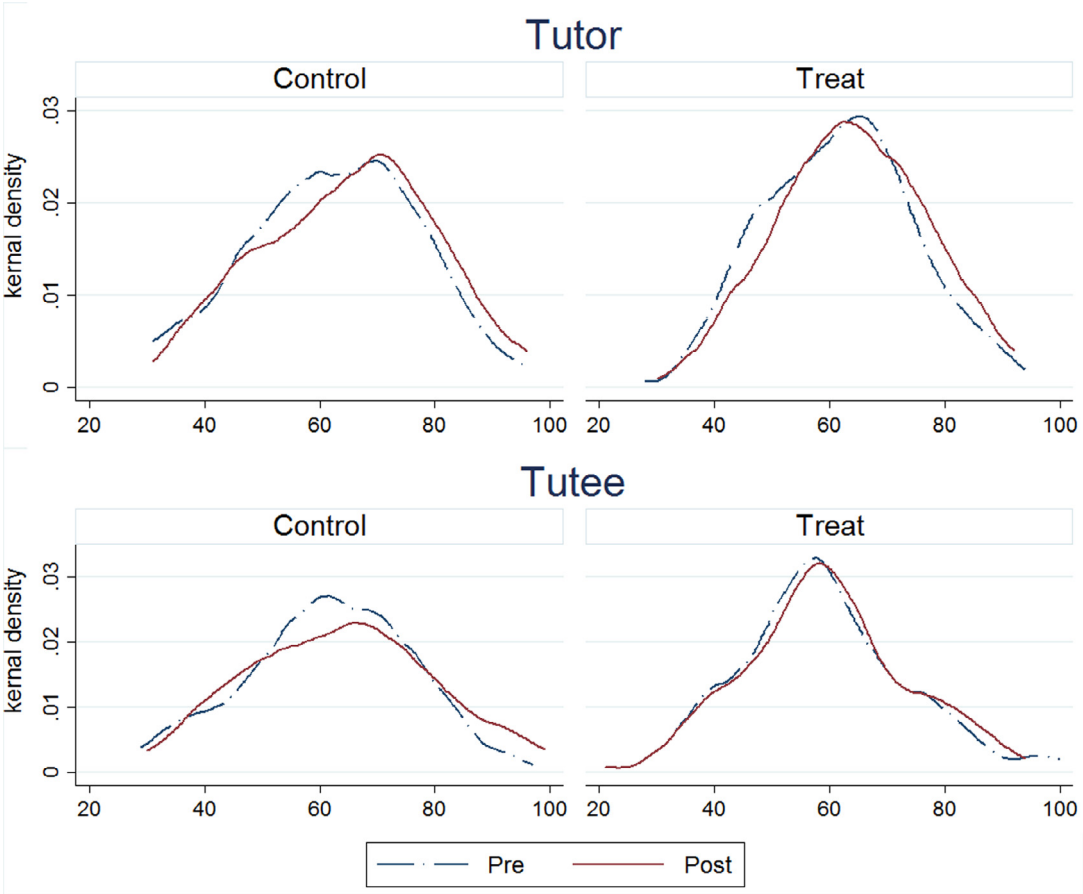

Fig. A2. Distributions of Mental Health Scores. Notes: Four plots show kernel density distributions of mental health scores for four different groups of students, divided by their tutor/tutee status and treatment/control assignment. Placebo tutor and tutee status in the control group are assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline.

and night self-study sessions, or during lunch and dinner breaks. Students were asked to take notes so that the class teacher could verify the interaction between the pair and reward them accordingly with snacks. The pairs who did not take notes or randomly copied something onto the notebook would not get a snack from the class teacher. Pairs who studied together four or more times each week received a bonus snack at the end of the week.

In addition, rewards were available to the student pairs based on the improvement in the average ranking of tutor and the tutee in the school's monthly exams that cover all subjects. We used a comprehensive exam so that students would not be incentivized to put more effort into only one subject (if performance on one were incentivized) at the expense of others. We chose to use the schools' own exams to encourage the students to closely follow and grasp what they were taught in school. This design also makes the incentivized peer tutoring program more scalable, because it does not require testing beyond what a school already conducts.

For midterm and final exams, students were rewarded according to the following procedure. First, we rank the pairs by improvement from their previous ranking, which is the average rank of both the tutor and tutee. Then we divided them into three tiers - a top third, middle third, and bottom third, with the top third consisting of most improved pairs. Each month, all students received a surprise prize, but the top third of improvers got a slightly superior prize to that of the middle third, and the middle third received a slightly better prize than the bottom third did. Prizes, such as notebooks, pens, or pencils, were diverse enough to differ in desirability. For example, if a tutor improved his ranking by 1 and a tutee improved his ranking by 3, the pair's average rank would increase by 2. If a first-tier pair's ranking improved by 2 to 5 ranks, this pair would get the best type of prize.

The outcome-based incentive scheme was designed to incorporate several features. First, the fact that everyone gets a prize, regardless of his or her performance, should maintain student participation and help foster a positive attitude toward the program. Second, since the reward is based on the improvement in ranking, all students, whatever their rank, have the opportunity to earn rewards. Third, it is unlikely that the same pair would improve or decline in average ranking each time, so different pairs are likely to get top prizes each month. However, if either member of the pair drops out, no reward will be given. This provides students with the incentive to provide social support to prevent their partner from dropping out of school.

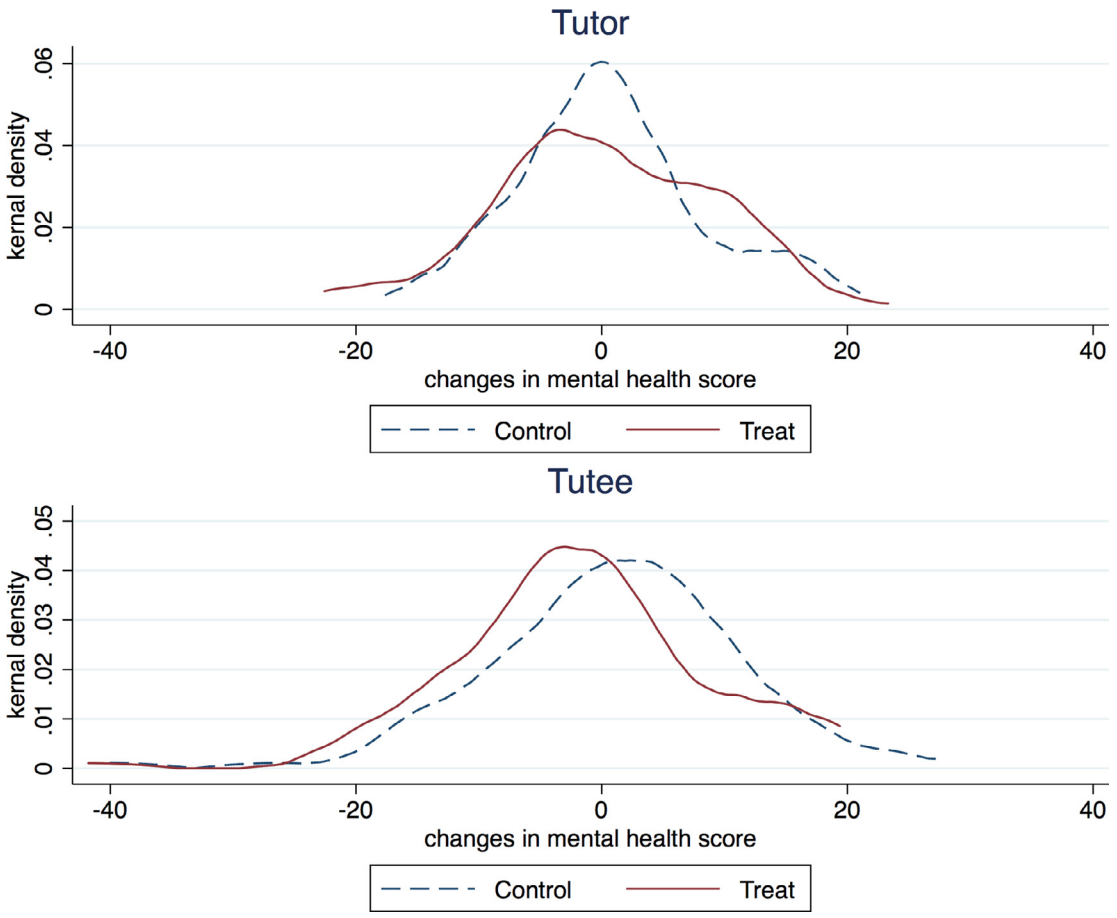

Fig. A3. Distributions of Change in Mental Health Scores. Notes: Each figure plots the residuals from regressing changes in mental health scores on baseline characteristics. The top figure only uses the tutor/ placebo tutor sample, and the bottom one only uses the tutee/placebo tutee sample. Placebo tutor and tutee status in the control group is assigned based on whether they are ranked in the top or lower half of class in terms of academic performance in the baseline. The p-values of Kolmogorov–Smirnov equality tests for tutor and tutee graphs are 0.565 and 0.017, respectively, indicating that tutees reported statistically significantly lower mental health scores.

Table A1
More balance check for tutor/tutee subsamples.

	Tutee			Tutor		
	Control	Treatment	Diff	Control	Treatment	Diff
family size	4.718	4.726	0.00786	4.818	4.907	0.0886
	(1.089)	(1.129)	(0.143)	(1.155)	(1.268)	(0.165)
boarding length (year)	2.479	2.507	0.0279	2.328	2.156	−0.172
	(2.286)	(2.120)	(0.285)	(2.129)	(1.974)	(0.281)
# boarding roommates	6.393	5.702	−0.692	5.556	5.377	−0.178
	(3.893)	(4.277)	(0.526)	(3.643)	(4.684)	(0.566)
class participation	5.017	3.605	**-1.412*** **	6.737	5.009	**−1.729***
	(4.657)	(2.686)	(0.494)	(7.731)	(7.146)	(1.020)
ask question	2.385	2.484	0.0993	2.475	2.432	−0.0425
	(0.539)	(0.518)	(0.0681)	(0.541)	(0.547)	(0.0741)
read	2.342	2.266	−0.0758	2.394	2.220	**−0.174** **
	(0.604)	(0.543)	(0.0741)	(0.586)	(0.525)	(0.0762)

(*continued on next page*)

Table A1 (*continued*)

	Tutee			Tutor		
	Control	Treatment	Diff	Control	Treatment	Diff
extra exercise	2.103	2	−0.103	2.111	2.153	0.0414
	(0.662)	(0.460)	(0.0738)	(0.604)	(0.500)	(0.0762)
bullied	1.615	1.669	0.0540	1.525	1.407	−0.118
	(0.808)	(0.707)	(0.0980)	(0.612)	(0.573)	(0.0810)
argument	3.325	1.847	**-1.478****	2.051	1.864	-0.186
	(6.270)	(2.157)	(0.611)	(2.106)	(3.197)	(0.363)
fight	1.302	0.952	-0.350	0.788	0.424	**-0.364***
	(2.964)	(1.701)	(0.315)	(1.774)	(0.964)	(0.199)
late homework	1.654	1.875	0.221	1.167	1.538	0.371
	(3.090)	(3.001)	(0.336)	(1.786)	(2.375)	(0.290)
late class	0.987	0.875	−0.112	1.071	0.996	−0.0749
	(1.822)	(1.561)	(0.210)	(1.736)	(1.353)	(0.195)
like school (1–10)	7.829	7.145	**-0.684***	7.414	7.585	0.171
	(2.922)	(3.017)	(0.383)	(2.657)	(2.292)	(0.340)

Note: Students with below-mean baseline math scores in the control classes consist of the control group for tutees in the treated classes. Similarly for tutors. The "Control" and "Treatment" columns show the means and standard deviations in parentheses. The "Diff" column shows the difference and robust standard errors in parentheses, with significance level indicated by $^{***}p< 0.01$, $^{**}p< 0.05$, and $^{*}p< 0.1$.

Table A2
Full specification of the main results.

Sample	(1) Tutors	(2)	(3) Tutees	(4)
Variables	std math	mental health score	std math	mental health score
treat	0.411*	−0.409	−0.0952	−4.115***
	(0.207)	(1.253)	(0.0858)	(1.279)
grade 7	−0.153	−2.030	−0.136	−1.892
	(0.231)	(1.454)	(0.0855)	(1.329)
female	−0.173	−1.141	−0.0522	0.383
	(0.145)	(1.822)	(0.120)	(1.084)
baseline math	0.305***	−0.157	0.375***	1.482**
	(0.0703)	(0.504)	(0.0575)	(0.655)
single parent	−0.0159	2.971**	0.543**	−2.028
	(0.188)	(1.250)	(0.194)	(2.194)
family size	−0.00798	−0.677	0.0218	0.124
	(0.0474)	(0.584)	(0.0432)	(0.851)
years of boarding school	−0.0424	0.0228	0.0239	−0.397
	(0.0515)	(0.525)	(0.0216)	(0.403)
# of boarding roommates	−0.00431	−0.228	−0.000309	−0.117
	(0.0199)	(0.254)	(0.0101)	(0.222)
baseline learningstress	−0.0425	−0.564***	−0.0515**	−0.417
	(0.0273)	(0.178)	(0.0232)	(0.336)
baseline social stress	0.0201	−0.344	−0.0285	−0.264
	(0.0205)	(0.393)	(0.0441)	(0.482)
baseline antisocial	0.0224	−0.266	−0.00922	−0.963
	(0.0329)	(0.338)	(0.0333)	(0.565)
baseline self-guilt	−0.0646*	−0.948**	0.0397	−0.960**
	(0.0343)	(0.397)	(0.0382)	(0.418)
baseline oversensitiveness	0.0679	−1.397***	0.0566**	−0.259
	(0.0620)	(0.299)	(0.0242)	(0.434)
baseline physical signs of stress	−0.0134	−1.298***	−0.00490	−1.293***
	(0.0326)	(0.172)	(0.0234)	(0.358)
baseline fear	0.00454	−0.301	0.00259	−1.491***
	(0.0321)	(0.346)	(0.0178)	(0.262)
baseline impulsiveness	−0.00242	−0.690*	0.0400	−0.291
	(0.0267)	(0.345)	(0.0409)	(0.591)
baseline like school	0.00212	−0.113	−0.0427*	−0.516**
	(0.0224)	(0.339)	(0.0219)	(0.178)
baseline class participation	0.00824	0.0470	0.0373***	−0.0899
	(0.0110)	(0.0646)	(0.0101)	(0.202)
baseline ask question	0.0935	−1.141	0.191	−2.816*
	(0.109)	(1.767)	(0.114)	(1.354)

(*continued on next page*)

Table A2 (*continued*)

Sample	(1) Tutors	(2)	(3) Tutees	(4)
Variables	std math	mental health score	std math	mental health score
baseline extra reading	−0.295**	−1.250	−0.00536	−1.191
	(0.113)	(1.570)	(0.104)	(2.029)
baseline extra practice problems	0.219	2.947**	-0.0940	1.597
	(0.151)	(1.039)	(0.137)	(2.178)
baseline class leader	−0.229*	−0.499	−0.296***	0.936
	(0.128)	(1.833)	(0.0909)	(1.777)
baseline bullied	−0.0409	−0.778	−0.175***	−1.383
	(0.0852)	(1.011)	(0.0561)	(1.315)
baseline argument	−0.00518	0.344	−0.00910	0.248*
	(0.0236)	(0.223)	(0.00823)	(0.121)
baseline fight	−0.0657*	−0.561	0.00487	−0.315
	(0.0342)	(0.381)	(0.0304)	(0.224)
baseline late hw	0.0339	−0.567	−0.0133	−0.280
	(0.0365)	(0.438)	(0.0163)	(0.306)
baseline late class	0.0377	−0.136	0.0339	−0.638
	(0.0500)	(0.466)	(0.0362)	(0.448)
baseline absence	−0.0112	0.213	−0.0440	−0.383
	(0.0518)	(0.758)	(0.0582)	(0.910)
Observations	208	198	223	211
R-squared	0.275	0.657	0.282	0.558

Robust standard errors in parentheses, clustered at the class level. Significance level: ***p< 0.01, **p< 0.05, *p< 0.1.

Table A3
Heterogeneous treatment effects on female students.

Depend. Var.	(1) std math score	(2)	(3) mental health score	(4)
Sample	Tutor	Tutee	Tutor	Tutee
treat	0.452	−0.0685	1.424	−3.846*
	(0.259)	(0.163)	(1.567)	(1.780)
female	−0.358*	0.0183	1.420	0.493
	(0.189)	(0.163)	(2.368)	(1.448)
treat X female	0.296	0.0282	−3.107	0.781
	(0.266)	(0.231)	(2.661)	(2.318)
Baseline Controls	Y	Y	Y	Y
Observations	197	209	190	200
R-squared	0.188	0.089	0.602	0.500

Robust standard errors in parentheses, clustered at the class level. ***p< 0.01, **p< 0.05, *p< 0.1.

Table A4
Student suggestions.

Category	Detail	# stu. response
Organization and arrangement		57
	more time/frequency	20
	more contact	10
	fix time of the day to do peer tutoring	8
	seat together	7
	more monitoring	5
	more friend conversations	3
	separate study & non-study students	2
	fix subject for each day	2
Matching mechanism		46
	More ppl in a group (esp. 4 ppl group)	22
	Self-choice pair	10
	Top-down pair	8
	Top-top pair	3
	Mix gender group	3
Incentive methods		18
	More prize for academic improvement/excellence	11
	No snack or change snack to other prizes (pen, book)	7
No Change/ No Suggestion		53

Note: In the evaluation survey, we also asked an open-ended question, "How do you think the peer tutoring program could be improved?" This table summarizes the frequencies of top suggestions.

References

Angrist, J., Lang, D., Oreopoulos, P., 2009. Incentives and services for college achievement: evidence from a randomized trial. Am. Econ. J. 1 (1), 136–163.
Blimpo, M.P., 2014. Team incentives for education in developing countries: a randomized field experiment in benin. Am. Econ. J. 6 (4), 90–109.
Burke, M.A., Sass, T.R., 2013. Classroom peer effects and student achievement. J. Labor Econ. 31 (1), 51–82.
Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Rev. Econ. Stat. 90 (3), 414–427.
Cohen, P.A., Kulik, J.A., Kulik, C.-L.C., 1982. Educational outcomes of tutoring: a meta-analysis of findings. Am. Educ. Res. J. 19 (2), 237–248.
Cornwell, C., Mustard, D.B., Sridhar, D.J., 2006. The enrollment effects of merit-based financial aid: evidence from Georgia's hope program. J. Labor Econ. 24 (4), 761–786.
Ding, W., Lehrer, S.F., 2007. Do peers affect student achievement in China's secondary schools? Rev. Econ. Stat. 89 (2), 300–312.
Fantuzzo, J.W., King, J.A., Heller, L.R., 1992. Effects of reciprocal peer tutoring on mathematics and school adjustment: a component analysis.. J. Educ. Psychol. 84 (3), 331.
Fryer, R.G., 2011. Financial incentives and student achievement: evidence from randomized trials. Q. J. Econ. 126 (4), 1755–1798.
Fuchs, D., Fuchs, L.S., Mathes, P.G., Simmons, D.C., 1997. Peer-assisted learning strategies: making classrooms more responsive to diversity. Am. Educ. Res. J. 34 (1), 174–206.
Gneezy, U., Meier, S., Rey-Biel, P., 2011. When and why incentives (don't) work to modify behavior. J. Econ. Perspectives 25 (4), 191–209.
Greenwood, C.R., Delquadri, J.C., Hall, R.V., 1989. Longitudinal effects of classwide peer tutoring.. J. Educ. Psychol. 81 (3), 371.
Hansen, B.B., Bowers, J., 2008. Covariate balance in simple, stratified and clustered comparative studies. Stat. Sci. 219–236.
Kang, C., 2007. Classroom peer effects and academic achievement: quasi-randomization evidence from south korea. J. Urban Econ. 61 (3), 458–495.
Kremer, M., Holla, A., 2009. Improving education in the developing world: what have we learned from randomized evaluations? Annu. Rev. Econ. 1, 513.
Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. Rev. Econ. Stat. 91 (3), 437–456.
Krueger, A.B., 1999. Experimental estimates of education production functions*. Q. J. Econ. 114 (2), 497–532.
Levin, H. M., et al., 1984. Cost-effectiveness of four educational interventions.
Li, T., Han, L., Zhang, L., Rozelle, S., 2014. Encouraging classroom peer interactions: evidence from chinese migrant schools. J. Public Econ. 111, 29–45.
Maher, C.A., Maher, B.C., Thurston, C.J., 1998. Disruptive students as tutors: a systems approach to planning and evaluation of programs. Peer-Assisted Learn. 145–163.
Rawlings, L.B., Rubio, G.M., 2005. Evaluating the impact of conditional cash transfer programs. World Bank Res. Obs. 20 (1), 29–55.
Rohrbeck, C.A., Ginsburg-Block, M.D., Fantuzzo, J.W., Miller, T.R., 2003. Peer-assisted learning interventions with elementary school students: a meta-analytic review.. J. Educ. Psychol. 95 (2), 240.
Sacerdote, B., 2011. Peer effects in education: how might they work, how big are they and how much do we know thus far? Handbook Econ. Educ. 3, 249–277.
Scruggs, T.E., Mastropieri, M.A., 1998. Tutoring and students with special needs. Peer-Assisted Learn. 165–182.
Sutherland, K.S., Wehby, J.H., Gunter, P.L., 2000. The effectiveness of cooperative learning with students with emotional and behavioral disorders: a literature review. Behav. Disord. 225–238.
Tolmie, A.K., Topping, K.J., Christie, D., Donaldson, C., Howe, C., Jessiman, E., Livingston, K., Thurston, A., 2010. Social effects of collaborative learning in primary schools. Learn. Instr. 20 (3), 177–191.
Topping, K.J., 2005. Trends in peer learning. Educ. Psychol. 25 (6), 631–645.
Webb, N.M., Farivar, S., 1994. Promoting helping behavior in cooperative small groups in middle school mathematics. Am. Educ. Res. J. 31 (2), 369–395.
Yi, H., Zhang, L., Luo, R., Shi, Y., Mo, D., Chen, X., Brinton, C., Rozelle, S., 2012. Dropping out: why are students leaving junior high in China's poor rural areas? Int. J. Educ. Dev. 32 (4), 555–563.