Teamwork and Human Capital Development*

Chunchao Wang Aiping Xiao Yu Zhou

Abstract:

This study investigates how teamwork influences students' human capital, which is defined to be academic performance and personality traits. In a rural county in China, we randomly select classes in elementary schools and form small teams within the treatment classes. Team members need to complete team activities. We find that the act of forming teams can significantly improve students' academic performance. Teamwork also causes substantial changes in noncognitive skills. Students in the treatment classes achieve higher scores in conscientiousness, extraversion, openness, and neuroticism but lower scores in agreeableness. These changes indicate a higher level of performance motivation.

Key words: Teamwork, human capital, academic performance, noncognitive skills, student behavior

JEL codes: C93, I21, J24, Z13

* Chunchao Wang is a professor of economics at Jinan University. Aiping Xiao is an assistant professor of economics at Guangdong University of Finance (Joint first author). Yu Zhou is an assistant professor of economics at Jinan University (corresponding author: zhouyu23@vt.edu). This work received financial support from the National Social Science Fund of China (18ZDA081), the Fundamental Research Funds for the Central Universities (19JNKY06), and the National Natural Science Foundation of China (72103050). The authors gratefully acknowledge the valuable comments and suggestions from Jorge Aguero, Te Bao, Xiqian Cai, Sudipta Sarangi, Robert Webb, Nick Webber, Jia Wu, Minqiang Zhao, and two anonymous referees. All errors remain our own. The experiment has been registered in AEA RCT system with the registration number AEARCTR-0008280. The data used in this article are available online: Chunchao Wang. Teamwork and Human Capital. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E153341V1

doi:10.3368/jhr.0121-11400R2

I. Introduction

Lack of motivation in students in basic education has been recognized as a worldwide phenomenon (Hanushek 2003; Glewwe and Muralidharan 2016; Muralidharan 2017; World Bank 2018). In the literature on peer effects, a trend shows that group-based incentives work better than individual-based incentives in improving students' academic performance (Blimpo 2014; Li et al. 2014).[1] However, the pecuniary incentives in the interventions may reduce the feasibility of implementation in developing areas. In addition, current interventions do not pay sufficient attention to noncognitive skills, which were found to be an important component of human capital.[2] Given that students often make prosocial decisions when faced with team incentives (Babcock et al. 2015), it is interesting to know whether the easily available intervention of teaming up students in school activities without pecuniary incentives could motivate students. Students' human capital development, which is defined as a combination of academic performance (cognitive skills) and personality traits (noncognitive skills) would also be affected by the intervention. In this study, we investigate how the human capital is affected by the teamwork. The experimental setting preserves the schools' teaching schedule and arrangement of activities. The class teachers, who is responsible for the class as a whole, report to the teams the team performance daily. The class

[1] Peer effects were extensively studied in the literature and were found in different stages of education and types of groups (Burke and Sass 2013; Jain and Kapoor 2015; Booij, Leuven, and Oosterbeek 2017; Feld and Zölitz 2017). Ding and Lehrer (2007), Carman and Zhang (2012), and Li et al. (2014) provided evidence on peer effects in Chinese high schools, middle schools, and elementary schools, respectively. Peer effects can be found in academic activities (Carrell, Fullerton, and West 2009; Lu and Anderson 2016; Li, Mak, and Wang 2019) and nonacademic activities (Lavy and Sand 2019). See Epple and Romano (2011) and Sacerdote (2011) for detailed reviews on peer effects.

[2] In a growing literature, noncognitive skills have become an important component of the definition of human capital explicitly (Currie and Almond 2011) and implicitly (Cunha and Heckman 2007).

teachers also keep a confidential record of each team member's performance. In an incentive-compatible environment, students need to cooperate with their teammates to outperform other teams. Perceived team incentives may be stronger in the teams than individual incentives, because ideally students need to consider the impact of their individual behaviors on the team. We track the changes in cognitive and noncognitive skills and explore the effects of teamwork on the students' human capital development fostered in our randomly formed teams. Beyond the mere quantification of the effects of teamwork, in this study we aim also to understand the underlying mechanism.

We randomly select classes in elementary schools as the treatment groups in a rural county in China and form small teams with five to six members within the classes. The students in the teams complete daily duty tasks, attendance checks, and counts of disruptive behaviors.[3] The students finish their homework individually but submit their homework as a team throughout an entire semester. The students in the control classes engage in the same tasks individually instead of as a team. Teachers provide reports on the students' disruptive behaviors and task performance in both types of classes. However, the reports on the treatment classes are based on the performance of the entire team. The students need to adapt themselves to the teamwork when completing the team tasks. Using the difference-in-difference (DID) method, we find that forming teams improves cognitive and noncognitive skills of the students in the treatment classes.[4]

[3] Daily duty tasks are greeting teachers and classmates in the morning, cleaning classrooms, and organizing exercises during class intermissions.

[4] The experimental design emphasized randomness as much as possible. However, individual and family characteristics, in which absolute randomness cannot be achieved have substantial effects on students' human capital. Compared with a direct between-group comparison, the DID method can help us properly control such factors.

Our experimental design and estimation strategy enable us to properly identify the effects of teamwork on human capital. First, the treatment classes are strictly selected via a lottery to minimize selection bias. The descriptive statistics indicate the absence of significant differences in prior academic performance and personality traits between the students in the treatment groups and control groups.[5] Second, the students in the control classes and treatment classes are required to take the same types and numbers of tasks. The only difference is that the students in the treatment classes have fixed partners in the team activities, such as daily duties. The experimental design minimizes interruption to the regular teaching schedule. The students in treatment classes are not required to stay in school longer or undertake more tasks than their counterparts in the control classes. Other factors that may potentially affect human capital are controlled, and any observed changes in human capital can be substantially attributed to the effects of teamwork. The mechanism analysis shows that changes in behaviors are the functioning mechanism behind the improvement in academic performance. The students changed their behaviors owing to self-policing, social stigma, and learning from role models. The setting of team appraisals as a reward functions as motivations to change personality traits.

This study contributes to the literature in three ways. First, we determine that teamwork elicits significant changes in the students' personality traits, and the changes reflect a high level of performance motivation in the students. Unlike pioneering

[5] In our sample, all the participating schools had classes in the treatment and control groups. The students in the treatment groups had slightly higher scores in neuroticism and lower scores in agreeableness than those in the control groups. These between-group differences increased with the introduction of forming teams; thus the initial difference in neuroticism and agreeableness did not bias our estimation on the other traits.

research on noncognitive skills that focused on the outcome of individual behaviors (for example, Evans, Oates, and Schwab 1992; Aizer 2008; Neidell and Waldfogel 2010) or mental stress and level of social acclimation and satisfaction in school (Gong, Lu, and Song 2021), we employ an alternative perspective that focuses on personality traits as direct measures of noncognitive skills (Heckman and Rubinstein 2001). Research on education provides evidence that the positive effects of teamwork can manifest in more than one dimension depending on the team activities.[6] In our experiment, we expect teamwork to induce changes beyond academic performance. We choose the "Big Five" personality traits to account for the complexity of the effects of teamwork. We find that the students in the treatment classes obtain high scores in conscientiousness, extraversion, openness, and neuroticism but low scores in agreeableness. All the changes in personality traits in our results indicate a high level of performance motivation (Judge and Ilies 2002; Hart et al. 2007).

Second, we establish that forming teams is a cost-effective way to improve education outcomes. Most of the related studies focus on how pecuniary team incentives can be used to improve teams' academic outcomes. Blimpo (2014) and Li et al. (2014) found that students in developing countries who received peer (team) incentives improved their academic performance.[7] Instead of focusing on the effects of pecuniary incentives, we identify the effects of teamwork alone. As pecuniary incentives are absent in our experiment, the students improve their performance solely

[6] Slavin (1980), Johnson and Johnson (2002), Hänze and Berger (2007), and Drakeford (2012), among other scholars, proposed that study teams can help improve the students' learning attitudes, school participation, academic performance, and social skills.

[7] In Li et al. (2014), the pecuniary incentive for teams is RMB 200 (approximately USD 29) per class. In the work of Blimpo (2014), the individual incentive varies from USD 10 to USD 30.

to achieve team success. The results of the heterogeneity tests suggest that teamwork exerts a large positive impact on the students in lower grade levels. Compared with pecuniary incentives, which may not be accessible to schools with constrained budgets, forming teams is more affordable and feasible.

Third, we provide new insights into the mechanisms that motivate students in teams. Our results suggest that the students in the teams have significantly fewer disruptive behaviors and better focus during lectures than their counterparts in the control classes. We further explore students' reasons for changing their behaviors by conducting a follow-up survey on all 15 class teachers who participated in the experiment and a 10% random sample of the students in the treatment classes. The results show that 93% of the teachers believe that giving the students information on their team performance and setting team appraisals as rewards are essential to motivate the students. The motivation is well-perceived as 90% of the students indicate that they paid more efforts to study and observe in-class discipline. The changes in personality traits also confirm the students' higher performance motivation level. In addition, the students report that they changed their behaviors owing to intrinsic and extrinsic reasons, such as self-policing, social stigma, and learning from role models. The improved academic performance is a result of the students' efforts to study.

In summary, the results of this study indicate that forming teams, by itself, can benefit students' human capital development. The effects of the intervention are reflected by not only the students' improved academic performance but also their personality traits, which may generate positive effects in the long run. Multiple

mechanisms exist behind the effectiveness of forming teams.

The remainder of this paper is organized as follows. Section II introduces the curriculum schedule and seating arrangement in elementary schools in China. Section III describes the experimental design and dataset. Section IV discusses the estimation strategy and empirical model. Sections V and VI present the empirical results and expound on the functioning mechanism of the teams, respectively. Finally, section VII concludes the paper.

II. Curriculum Schedule and Seating Arrangement in Elementary Schools in Rural China

Elementary school students in rural China are enrolled in units of classes. Once students are enrolled in a school and assigned to a class, they typically remain in the same class for the entirety of their six-year primary education. Class transfers are possible but rare.[8] On weekdays, students generally arrive at school at 8:00 a.m. and leave at 4:30 p.m. They attend six lectures with a duration of 45 minutes, with a 10-minute intermission between lectures. Many students spend lunch breaks in class. The curriculum schedule ensures that students spend most of their school hours with their classmates. Outside of school, apart from doing homework, students spend time on activities such as extracurricular reading or chores.[9]

The seating arrangement in elementary schools in rural China is relatively fixed.

[8] We did not observe any class transfers during the experiment period.
[9] After-school classes are uncommon in rural China. In our sample, fewer than 1% (11 out of 1,589) of the students were enrolled in after-school classes.

Students are assigned to different rows according to their height, that is, short students sit in the front rows for the practical reason that tall students may block the view of short students if they sat in the front rows. Shuffling is implemented to help the students' vision development but occurs mainly by columns instead of rows. Students generally sit in pairs, and seatmates share a desk or have individual but connected desks. Seatmate pairings remain fixed even with shuffling.

The fixed seating arrangement and curriculum schedule in elementary schools in China provide us with ample opportunities to conduct the field experiment. We randomly assigned students with similar heights to teams and ensured that the team members sat next to one another in either the same row or same column.[10] We expected physical proximity to increase the students' awareness of the teams with minimum interruption, if any, to the common seating arrangement. Applying the seating arrangement rule to both types of classes meant that the students in the treatment classes would have sat in the same position and have had the same peers if they were in the control classes. The difference, if any, between the treatment classes and control classes can be attributed to forming teams.

[10] Section III provides detailed information on the seating arrangement. We are aware of the potential endogeneity between the students' academic performance and height (Case and Paxson 2008; Vogl 2014) and thus controlled for the students' height in the analysis. We found that possible endogeneity did not bias our results. Another potential endogeneity is that people may feel more comfortable communicating with people with a similar height. However, as the seating arrangement in the control classes was also based on height, the students in the treatment classes would have been assigned to the same rows if they were in the control classes. As the question in this study is whether teams are effective in improving students' human capital, the current experimental design can properly identify the effects of teamwork, because the same seating rules were applied to the treatment classes and control classes.

III. Experimental Design

We conducted the experiment intervention in a rural county (L County) of Hunan Province in Central China for five months, from September 2015 to January 2016. The duration of the intervention covered the entire autumn semester of the academic year 2015/2016.[11] In addition, we invited all 15 participating class teachers and a 10% random sample of students in the treatment classes to answer a follow-up survey in the spring semester 2021, five years after the end of the intervention. With the permission and cooperation of the local education bureau, we randomly selected five elementary schools from the complete list of elementary schools in the county to conduct the experiment.[12] The students in the participating schools were initially assigned to different classes randomly at the start of first grade. We focused on the students in the third, fourth, and fifth grades.[13] We randomly chose two classes in each grade from each school. Via a lottery, we assigned one class as a treatment class, wherein we randomly formed teams. We assigned the other class as the control class, wherein no intervention was implemented. In this section, we describe how the teams were formed. In total, we had 30 classes comprising 15 treatment classes and 15 control classes. A total of 1,589 students made up the 30 classes, specifically 907 male students (57%), and 682 female students (43%). Table 1 presents the gender composition of the sample schools.[14]

[11] In Chinese elementary schools, the autumn semester starts on September 1 and ends around January 15.
[12] Online Appendix A provides the cooperation agreement. Online appendix is available to be downloaded at https://doi.org/10.3886/E153341V1
[13] We excluded first- and second-grade students, because their literacy and comprehension capabilities may be insufficient for the completion of the questionnaires. We also excluded sixth-grade students, because they were in transition to middle school, which could cause difficulties in follow-up survey.
[14] Male students were overrepresented in our sample for two possible reasons. First, the county where we conducted the experiment had an extremely high percentage of people working as migrant workers in urban areas.

To measure academic performance, we collected information on the students' scores from three examinations, that is, their final examination in the spring semester of academic year 2014/2015 and the midterm examination and final examination in the autumn semester of academic year 2015/2016. We used the scores from the spring semester as the baseline measurement, because they reflected the academic performance of the students before their participation in the experiment. For each examination, scores from three subjects, Chinese, mathematics, and English, were reported.[15] All examinations were unified and designed by the local education bureau, and all the participating schools administered the examinations at the same time. Thus, the scores were comparable across schools.

During the intervention, we requested all the students answer a paper-based questionnaire twice. Online Appendix B presents the invitation letter and instructions for the students. The first-round questionnaire was given two weeks before the start of the autumn semester, and the second-round questionnaire was administered two weeks before the final examination. The questionnaire collected information on the students' demographics and attitudes toward studying and personality traits measurements. The questionnaire also gathered basic information on the students' parents, such as their education, occupation, and income. The students answered the questionnaires during self-study sessions in the presence of class teachers, who then collected the forms and

As parents move to urban areas, they are likely to bring their daughters with them for safety concerns. Therefore, boys may be overrepresented among the children left behind. Second, biased birth gender favoritism is more severe in rural areas than in urban areas. Such favoritism would result in a higher percentage of boys in the population.

[15] Chinese, mathematics, and English are the three main subjects in Chinese elementary schools. The students in the participating schools do not study English before third grade; thus, we exclude the English scores from the baseline measurement for the third-grade students.

uploaded the detailed information to our online system. To guarantee data accuracy and completeness, the class teachers returned incomplete forms or forms with obvious errors to the corresponding students for correction until the forms were satisfactory. We supervised the entire procedure of questionnaire completion and uploading to ensure the accuracy and effectiveness of the information.

In order to understand the mechanisms behind the effects of forming teams, we also conducted a follow-up survey on the class teachers and a random sample of the students in treatment classes. In the teacher's survey, we asked the class teachers about their observations on the changes in the students' behaviors and the general experimental implementation. In the student's survey, we asked the students to reflect on their behavioral changes during the experiment. The questions were designed to explore the reasons behind their behavioral changes. We provide the details of the follow-up survey in section VI.

Forming teams and arranging seats were crucial components of the experiment. We set the team size to five to six members, which is within the range of the optimal team size noted in the literature (Drakeford 2012). We justify the practice of forming teams as follows. In Chinese classrooms, students are seated in rows from the front to the back of the classroom, with short students taking the front rows. To be comparable with the current classroom structure, we first assigned the students to three sets according to height, that is, below average, average, and above average height. In each set, every six students were grouped by lottery, and their seats were also assigned by

lottery.[16] We ensured that the short students sat in the front rows, which was how they would be assigned if they were in the control classes. We also ensured that the members of a same team sat next to one another either in the same row or in the same column. In consideration of the students' visual development, seating was shuffled every two weeks, but the team members remained unchanged. Figure 1 illustrates the seating arrangement in the experiment classes.

We selected some standard school activities as team tasks to increase interaction within the teams and raise team awareness. Such tasks included daily duties, attendance checks, disruptive behavior warnings as a team, and homework as a team. Homework was completed individually, but the team members were required to submit their homework on time as a team.[17] In the schools in our experiment, examination grades were determined by only the students' performance in examinations. Homework grades were not a component of the final grade. The academic performance in the analysis was measured by the standardized examination scores. Students took the examinations on their own. The scores reflected academic ability at the individual level. The class teachers reported the team performance daily to the class. The records included the total number of disruptive behaviors of each team and the team performance on the team tasks at the aggregate level. A confidential record of disruptive behaviors at the individual level was also kept by class teachers. This record was used in the mechanism

[16] Tall students may block the view of short students sitting behind them. To solve this problem, we teamed up the students according to height. This arrangement may sacrifice a certain level of randomness in the seating arrangement but was a practical necessity. To minimize the potential endogeneity between the students' height and academic achievement, we controlled for seating arrangement and height in the empirical analysis.

[17] Although team activities, such as cooperative learning, which requires students to undertake a certain level of teaching, can increase interaction among students, such activities can also alter the teachers' behaviors in the experimental classes. To identify the specific effects of teams, we limited the team activities to those that did not alter the teachers' behaviors.

analysis of Section VI.

Teams may have differentiated effects on students in the geographic core of a team and those on the periphery owing to different exposures to members of other teams. However, the seating arrangement and seating shuffling in our experiment ensured that the students, except those sitting in the first and last rows, sat in the core and periphery of their team for approximately the same length of time.

In the control classes, the seating arrangement was also set according to the students' height and by lottery without forming teams. The similar seating arrangement in both types of classes implied that the students in the treatment classes would have sat at the same position if they were in the control classes. After setting the seating arrangement, we instructed the class teachers to inform us of any changes. However, no changes in the seating arrangement were reported. The students in the control classes also received a record of their individual in-class behaviors and school task performance. Figure 2 illustrates the experimental procedure.

A feature of our sample is that owing to limited educational resources, the same set of teachers, class teachers, and subject teachers was assigned to the treatment and control classes in the same grade in each school. The teachers received training before the experiment implementation. We specifically requested the teachers to treat the control classes and treatment classes equally and to form teams only in the treatment classes. We expected the training to control for the teaching quality and prevent potential spillover effects. In our follow-up survey, all the class teachers indicated that they did not treat the treatment classes differently such as by paying more attention, or

by using a different curriculum or pedagogy. The only exception was that class teachers reported the students' performance based on team performance in the treatment classes and based on individual performance in the control classes. Yet, class teachers kept a record of disruptive behaviors in both treatment classes and control classes. The teachers did not receive any pecuniary incentives from this experiment based on the students' academic performance. We also requested the inspectors, one for each grade in a school, to observe and ensure that the teachers did not treat the control and treatment classes differently. This feature of our methodology eased concerns that the experiment might cause changes in behaviors of the teachers.

In summary, owing to our sample selection strategy, the treatment classes and control classes were completely comparable. In the treatment classes, forming teams would raise the students' awareness of teamwork. Their nearby classmates would have been the same ones if they were in the control classes. The students undertook the same number and types of activities on school days. Finally, we expect the teachers' quality and behaviors to be the same in both groups of classes.

IV. Empirical Model

We employed the DID method to estimate the effects of teamwork on academic performance (cognitive skills) and personality traits (noncognitive skills). The experimental design emphasized randomness to the highest potential. However, individual and family characteristics could present considerable variation at the individual level and influence students' human capital substantially. The DID analysis

can properly control for factors external to the teams that could affect the students' cognitive and noncognitive skills. The validity of the DID analysis relied on the randomness of the treatment classes and teams. To avoid endogeneity, we enforced randomness in choosing the treatment classes and forming the teams. We conducted a between-group t-test on the means of the baseline scores, noncognitive skills, and individual characteristics from the first-round questionnaire. Table 2 reports the descriptive statistics. The results showed no significant differences in the baseline scores. No significant differences were observed in the personality traits of openness, conscientiousness, and extraversion and in terms of gender, age, and height. Significant differences were found only in the agreeableness and neuroticism dimensions of the "Big Five" personality traits. We regarded the differences in the two dimensions as a random fluctuation. In addition, our empirical analysis revealed that these initial differences did not bias our results.

We controlled for several sets of variables that influenced academic performance and personality traits. Individual characteristics were gender, age, and height. Class characteristics were school, grade, class, and team dummy variables. Family characteristics were parents' income, extracurricular reading time, and household chore time. We also employed the baseline examination scores to control for the students' initial study abilities. Our model is presented as follows:

$$y_{it} = \beta_0 + \gamma treatment_i \times post_t + \beta_1 treatment_i + \beta_2 post_t + \beta_3 baseline_i + \boldsymbol{X_i}\boldsymbol{\Theta} + \epsilon_{it}, \quad (1)$$

Where y_{it} denotes the standardized examination scores or noncognitive skills of

student i at time t; $\boldsymbol{X_i}$ is the vector of individual characteristics, class characteristics, and family characteristics; $treatment_i$ represents the dummy variable "treatment classes," where 1 = treatment classes, and 0 = control classes; and $post_t$ is the time dummy, where 1 = experiment implementation, and 0 = pre-implementation. β_1 is the unobserved fixed effects in the treatment classes, β_2 pertains to the unobserved time effects in the treatment and control classes, β_3 refers to the students' initial study abilities, ϵ_{it} denotes the error term, and γ is the coefficient on the cross-term between the treatment and post dummies. γ is the core coefficient in our research and presents the effects of teamwork.

V. Empirical Results

In this section, we present the descriptive statistics of the key variables, the main results of the effects of teams, in particular, the effects of teams on cognitive and noncognitive skills, and the heterogeneity tests on the effects of teams on academic performance.

A. Descriptive Statistics of Variables

Table 3 provides the descriptive statistics of the dependent and independent variables. We measured cognitive skills using standardized scores. Standardization was applied to the average scores in Chinese, mathematics, and English to yield standardized scores with a mean of 0 and a standard deviation (s.d.) of 1.[18]

[18] We used three standardized scores in our analysis, that is, standardized final exam scores, standardized midterm exam scores, and standardized baseline scores. For each examination, we first took the average on the scores of three subjects. Then the standardization was implemented on the averaged scores. Only the scores in Chinese and mathematics were used to construct the baseline scores of the third grade.

Noncognitive skills were represented by the "Big Five" personality traits, namely, openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

B. Effects of Teamwork on Cognitive Skills

Table 4 reports the effects of forming teams on cognitive skills. As the treatment classes and control classes were paired at each grade in a school, we clustered the standard errors at the school-grade level to control for the "paired" feature (de Chaisemartin and Ramirez-Cuellar 2020). In panel A, we tested whether the treatment class dummy was correlated with the students' test scores by using either the midterm examination scores or final examination scores as the dependent variables in an OLS regression. We added the individual characteristics, school characteristics, and family characteristics to the regression by steps. We observed a robust correlation between the treatment dummy and the dependent variables. In panel B, we used DID analysis on Equation (1) to test the effects of forming teams on the test scores. We added the control variables by steps. Columns (1) to (4) display the results of the comparison between the midterm examination scores and baseline examination scores with different model specifications.[19] The students took their midterm examinations two months after the start of the experiment. Forming teams significantly improved the students' academic performance in their midterm examination. On average, being in a team increased a student's standardized score by 0.090–0.100 s.d., depending on the model specifications. All the results were significant at the 1% level. Columns (5) to (8)

[19] As we did not ask the students to answer the questionnaire before their midterm examinations to update their individual and family characteristics, we used the information collected in the first-round questionnaire.

present the results of the DID analysis comparing the final examination scores and baseline scores, which were taken five months after the start of the experiment. Compared with the effects observed in the midterm examination, the teams exerted a slightly weaker effect on improving the students' academic performance in the final examination. With the full model specification, we observe that the students' scores increased by 0.085 s.d. on average from the baseline examination to the final examination. In the last row of panel B, we conducted a small-sample inference, and the results confirmed the robustness of the results.[20]

C. Heterogeneity Tests on Cognitive Skills

We searched for evidence of the various effects of teamwork across the individual characteristics (i.e., gender, academic performance, and grade). Table 5 summarizes the heterogeneous effects.

The first two columns in Table 5 test the gender differences in the effects of teamwork on academic performance. No significant gender differences existed in our results.

We also explored whether the teams exerted different effects on the students with different academic abilities. We divided the students' baseline scores into four tiers, that is, A (above 85), B (75–84), C (65–74), and D (below 64). Columns (3) to (6) in Table 5 report the effects for each tier. The results indicated that forming teams had a positive effect on all the tiers but was not significant for tiers A, B, or C. The between-group

[20] Finite sample inference *p*-values were computed with the Stata command *ritest* (Hess 2017). We took 1,000 permutations on the variable "$treatment \times post$". The permutation was implemented at the school-grade level in order to keep the pairwise data structure.

tests suggested that no significant differences existed in forming teams for the students with different academic abilities.

Students in lower grades are typically less mature than students in higher grades in terms of cognitive development. Students in lower grades may be more malleable than students in higher grades when the new concept of teams is introduced. We hypothesized that the students in lower grades can benefit more from teams compared with those in the higher grades. Columns (7) to (9) in Table 5 report the effects of teams on the third-, fourth-, and fifth-grade students, respectively. The effects of forming teams on the third grade were the largest, with scores increasing by 0.186 and 0.165 s.d. in the midterm and final examinations, respectively. Both results are significant at the 1% level and thus supported our hypothesis.

D. Effects of Teamwork on Noncognitive Skills

In this section, we investigate how forming teams affected the students' personality traits. We intentionally designed team tasks to raise the students' awareness of teamwork. Frequent communication and interaction were observed among the students when carrying out their team tasks. We expected the students' personalities to be influenced and shaped by the interactions. Eagerness for positive appraisal may also elicit motivation, which can be reflected in the changes in the students' personality traits. Our questionnaire probed the students' attitudes through several statements from which the "Big Five" personality traits were constructed (Goldberg 1990, 1992). We standardized the personality traits to have a mean of 0 and a s.d. of 1. Table 6 presents the differences in personality traits between the treatment and control classes.

The results showed the significant effects of teams on the students' personality traits. Of the five dimensions, conscientiousness and extraversion presented the most prominent differences in scores. Compared the results from two rounds of questionnaire, being a member of a team provided the students in the treatment classes with higher scores than their counterparts in the control classes. The between-group difference increased by 0.194 and 0.153 s.d. in conscientiousness and extraversion, respectively. The scores in openness also increased in the treatment classes, though the between-group differences were not significant. Being in a team slightly decreased agreeableness in the students by 0.042 s.d. and increased neuroticism by 0.072 s.d.

In the literature on psychology, the correlation between personality traits and motivation has been extensively studied.[21] Conscientiousness was found to be the strongest and most consistent trait correlated with performance motivation. Neuroticism and extraversion were also found to be positively correlated with motivation. Hart et al. (2007) distinguished intrinsic achievement motivation from extrinsic motivation. Conscientiousness, openness, and extraversion were found to be positively correlated with the intrinsic achievement motivation, whereas conscientiousness and neuroticism were positively related to the extrinsic achievement motivation. Agreeableness was also found to be negatively associated with extrinsic achievement motivation. Our results match these findings and showed that forming teams provided the students with incentives to perform well in school. The high scores in conscientiousness and neuroticism and low scores in agreeableness implied that

[21] See Judge and Ilies (2002) for a meta-analysis.

eagerness for positive appraisal created extrinsic motivation for the students. Moreover, evidence showed that the students internalized some motivations because their scores in extraversion also increased. The high motivation level may exert a large effect on the students' labor market outcome in the long run (Heckman and Mosso 2014).

To identify the effects of teamwork on the students' noncognitive skills, we performed a DID analysis on the noncognitive skills, as shown in Equation (1). Table 7 presents the results of the OLS estimates and the DID estimates. Forming teams increased conscientiousness and neuroticism by 0.140 and 0.200 s.d., respectively, and decreased agreeableness by 0.149 s.d. The effects are comparable in size to those reported in the literature of intervention on personality (Roberts et al. 2017). These outcomes suggested that being a member of a team raised the students' awareness of teamwork and responsibility, and eagerness for positive appraisal served as a motivation to perform well.

We also conducted heterogeneity tests on gender, academic performance, and grade levels. The results are summarized in Table 8. The boys and students with the lowest initial academic performance had a large increment of scores in conscientiousness and extraversion. The students in the lower grades achieved high scores in conscientiousness. The girls and students in the higher grades received low scores in agreeableness and high scores in neuroticism without significant changes in either conscientiousness or extraversion. The between-group tests suggested no significant differences between the different cohorts except that the students in the higher grade demonstrated low agreeableness and high neuroticism. The outcomes

suggested that forming teams created incentives for all the students regardless of gender, academic performance, and grade.

VI. Mechanism Analysis

In this section, we search for the potential mechanisms behind forming teams. As motivation should be reflected directly by the changes in behaviors, first, we analyzed the data on the students' in-class behaviors collected during the experiment intervention. Second, we analyzed the data from the follow-up surveys on the teachers and students about their perception and comments on the working mechanisms.

We focused on two types of in-class behaviors, namely, class discipline and in-class attention. Class discipline has an effect on study outcomes. Disruptive behaviors such as talking without permission and pranking can not only distract students' attention but also cause negative externalities on their neighbors and eventually lead to poor academic performance. We collected the students' disruptive behaviors reported by their teachers. The disruptive behaviors were talking without permission, pranking, reading comic books in class, sleeping in class, and tardiness (late to school). All the behaviors were measured as frequency per week. In-class attention correlates with class discipline but differs by measuring the effort exerted by a student to study. For example, daydreaming students may not demonstrate any disruptive behavior, but their lack of focus hinders the effectiveness of their study. We collected self-reported in-class attention levels from the students. The students reported their attention level as "mostly cannot pay attention," "not sure," or "mostly can pay attention."

Table 9 shows a between-group comparison of class discipline. Initially, the students in the treatment classes had low frequencies of talking without permission and pranking but also exhibited high frequencies of reading comic books, sleeping in class, and tardiness. No significant between-group differences in in-class attention were observed. In the second-round questionnaire, the between-group t-tests indicated that the students in the treatment classes improved their in-class behaviors in all dimensions, because they engaged in fewer disruptive behaviors than the students in control classes did. The between-group differences were significant in talking without permission, pranking, and reading comic books. Although the between-group differences were not significant in sleeping in class and tardiness, the students in teams had fewer faults than their counterparts in the control classes. Interestingly, the frequency of disruptive behaviors increased in the treatment classes and control classes. However, the frequency increased faster in the control classes than in the treatment classes. Forming teams seemed to prevent class discipline from deteriorating. The results also showed that the students in the treatment classes had significantly higher attention levels than their counterparts in the control classes.

We conducted a DID analysis of the students' behavior. Class teachers of both the control classes and treatment classes kept a record of the students' behavior at the individual level. As we controlled for the changes in the other factors influencing the students' behaviors such as individual characteristics and family characteristics, we were able to single out the changes in behaviors induced by teams. The results are reported in Table 10. When we controlled for the other factors, we found that being a

member of a team reduced the frequencies of talking without permission, pranking, and reading comic books by 1.96, 1.56, and 0.80 times per week, respectively, which accounted for 58%, 49%, and 81% of the corresponding disruptive behaviors reported in the second-round questionnaire. Joining a team also increased the students' in-class attention significantly. The improved academic performance was a result of improved in-class behaviors, as predicted in the literature on education (Wentzel 1991; Malecki and Elliot 2002).

We conducted a follow-up survey in the experiment location in spring semester 2021. The survey covered the full sample of 15 class teachers and a 10% random sample (80 students) of the students in the treatment classes.[22] The randomness of the survey sample was confirmed by the comparison with the treatment sample in Table 11.[23] The survey questionnaire is presented in the Appendix. In the survey on the class teachers, we asked seven questions about their observations on the changes in the students' behaviors. In addition, four questions were about the general experiment implementation. On a scale of 1 to 5 representing "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree," the teachers indicated whether they agreed with the statements. The results are reported in Table 12. All 15 teachers indicated that they could recollect the experiment, and 93% confirmed the strong connections between the team members (question 1), and 87% observed mutual supervision between the students on studying and in-class discipline (questions 2 and 3). The effectiveness of the teams was also confirmed by the teachers as 80% believed

[22] All 15 class teachers answered the survey and 73 out of 80 students (attrition rate 9%) answered the survey.
[23] Table A1 provides additional evidence on the randomness of the survey sample.

that forming teams improved the students' test scores (question 6).

For the open questions, we invited the teachers to comment on the mechanism that made the teams effective, and 93% of the teachers believed that giving students the information on their team performance and setting team appraisals as a common team goal were essential to motivate the students. The teachers also observed intensified competition between the teams to receive positive team appraisals. Moreover, the teachers mentioned that the mechanisms listed in the questionnaire such as satisfactory in-class discipline and mutual help, could be helpful in improving the students' academic performance. A total of 10 teachers (67% of the sample) mentioned that they appreciated the opportunity to join the experiment and adopted the student teams for all their classes after the experiment ended.

In the follow-up survey conducted on the students, we asked them to reflect on their behavioral changes during the experiment. The questions were designed to explore the reasons behind the changes. Specifically, we wanted to determine whether the changes occurred owing to self-policing or social stigma to misbehaving or learning from role models. If the students answered that they changed their behavior because of pressure from other team members from misbehaving (questions 11 and 12), then social stigma was considered as the reason for the behavioral change. If the students answered that they changed their behavior voluntarily (questions 9 and 10), then self-policing was considered as the reason for the behavioral change. Learning from role models was another potential reasons investigated (questions 13 and 14). For the statements, students were asked to give their opinion using a scale ranging from 1 to 5 representing

"strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree," respectively. The survey results of the students are reported in Table 13.

In general, joining teams had a positive effect on the students as 93% considered themselves more responsible (question 4), and 92% considered themselves more cooperative after joining a team (question 15). The student teams also cultivated an environment for the students to make friends (question 7). Forming connections with the other team members took time but the connections were strong once they were formed (questions 1, 2, and 9). In addition, 90% of the students indicated that they exerted considerable efforts to study and observe in-class discipline to obtain positive team appraisals (question 6). The results matched the class teachers' observations that the students were motivated to study and behave in class. The high level of performance motivation was reflected by the changes in the students' personality traits (Judge and Ilies 2002). We tried to distinguish self-policing from social stigma and learning from role models. Though we received more confirmative answers (strongly agree or agree) on self-policing and learning from role models than on social stigma, the Kolmogorov-Smirnov tests in Table 13 showed that all three reasons were equally important in changing the students' behaviors, as the answers were from the same distribution.

The follow-up surveys confirmed that setting team appraisals as a common team goal functioned as the mechanism to elicit the students' efforts. Owing to their motivation, the students' personality traits changed, and they exerted considerable efforts to study and observe in-class discipline. In psychology theory of group effort, a trigger of an individual's effort is a shared highly-valued team goal (Fishbach et al.

2016). Our observation had a good fit in the theory.

Another potential mechanism that induced changes in academic performance was changes in noncognitive skills. We provide the evidence for the strong correlation between noncognitive skills and cognitive skills in Table 14. In Table 15, we regress the test scores on the treatment effect by controlling for the noncognitive skills. The treatment effect remained significant, which implied the existence of additional mechanisms influencing cognitive skills other than changes in noncognitive skills. In addition, as we observed simultaneous changes in the cognitive and noncognitive skills, determining the causality in our current study was not feasible. Nevertheless, noncognitive skills are a potential mechanism to explore.

In summary, setting team appraisals as a common team goal was the mechanism that motivated the students. The students' personality traits and in-class behaviors changed as they became motivated. The improved in-class behaviors eventually resulted in enhanced academic performance in the treatment classes.

VII. Conclusion

In this study, we investigate the effects of teamwork on the human capital development of elementary school students. The randomization in the field experiment enables us to properly identify the effects of teamwork. Furthermore, we conduct a follow-up survey to identify the important mechanism behind forming teams.

We find that forming teams, by itself, increases the students' academic performance by 0.085–0.100 s.d., which is salient in the students in lower grades. With

respect to noncognitive skills, joining teams increases the students' scores in conscientiousness, extraversion, and neuroticism and lowers their scores in agreeableness. No significant change is observed in the students' openness. Based on the literature on psychology as a reference, the results suggest that forming teams provides the students with motivation to perform well owing to their eagerness to receive positive team appraisals. In the mechanism analysis, we provide evidence that the students in the treatment classes demonstrate better in-class behaviors than their counterparts in the control classes. The improved academic performance is a result of the efforts to study. In the follow-up survey, 93% of the teachers believe that providing feedback to the students and setting team appraisals as a common team goal are essential to motivate students. In addition, 90% of the students indicate that they exerted considerable efforts to study and observe in-class discipline to receive positive team appraisals. The survey results show that changing behaviors voluntarily (self-policing), pressure from teammates (social stigma), and learning from role models are the reasons behind the students' behavioral changes.

Overall, forming teams is a cost-effective way to provide incentives that can be easily scaled up in resource-strained environments. Human capital, that is, cognitive and noncognitive skills, can be improved through teamwork. In the future study, exploring the effective ways to organize teams would be interesting.

Multiple extensions can be conducted based on our current findings. Determining the effects of including academic activities, such as cooperative learning, on the effectiveness of student teams would also be interesting. Investigations on the

mechanisms, such as determining the influence of changes in noncognitive skills on changes in cognitive skills, can provide a new avenue for conducting behavior intervention.

References

Aizer, Anna. 2008. “Peer Effects and Human Capital Accumulation: The Externalities of ADD.” *NBER Working Paper* w14354.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2015. “Letting Down the Team? Social Effects of Team Incentives.” *Journal of the European Economic Association* 13(5): 841–870.

Blimpo, Moussa P. 2014. “Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin.” *American Economic Journal: Applied Economics* 6(4): 90–109.

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. 2017. “Ability Peer Effects in University: Evidence from a Randomized Experiment.” *Review of Economic Studies* 84(2): 547–578.

Burke, Mary A., and Tim R. Sass. 2013. “Classroom Peer Effects and Student Achievement.” *Journal of Labor Economics* 31(1): 51–82.

Carman, Katherine Grace, and Lei Zhang. 2012. “Classroom Peer Effects and Academic Achievement: Evidence from a Chinese Middle School.” *China Economic Review* 23(2): 223–237.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. “Does Your Cohort Matter? Measuring Peer Effects in College Achievement.” *Journal of Labor Economics* 27(3): 439–464.

Case, Anne, and Christina Paxson. 2008. “Stature and Status: Height, Ability, and Labor Market Outcomes.” *Journal of Political Economy* 116(3): 499–532.

Cunha, Flavio, and James J. Heckman. 2007. “The Technology of Skill Formation.” *American Economic Review* 97(2): 31–47.

Currie, Janet, and Douglas Almond. 2011. “Human Capital Development before Age Five.” In *Handbook of Labor Economics* Vol. 4, ed. David Card and Orley Ashenfelter, 1315–1486. North-Holland: Elsevier.

de Chaisemartin, Clément, and Jaime Ramirez-Cuellar. 2020. “At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?” *NBER Working Paper* w27609.

Ding, Weili, and Steven F. Lehrer. 2007. “Do Peers Affect Student Achievement in China’s Secondary Schools?” *Review of Economics and Statistics* 89(2): 300–312.

Drakeford, William. 2012. “The Effects of Cooperative Learning on the Classroom Participation of Students Placed at Risk for Societal Failure.” *Psychology Research* 2(4): 239–246.

Epple, Dennis, and Richard E. Romano. 2011. “Peer Effects in Education: A Survey of the Theory and Evidence.” In *Handbook of Social Economics* Vol. 1, ed. Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, 1053–1163. North-Holland: Elsevier.

Evans, William N., Wallace E. Oates, and Robert M. Schwab. 1992. “Measuring Peer Group Effects: A Study of Teenage Behavior.” *Journal of Political Economy* 100(5): 966–1991.

Feld, Jan, and Ulf Zölitz. 2017. “Understanding Peer Effects: On the Nature, Estimation,

and Channels of Peer Effects." *Journal of Labor Economics* 35(2): 387–428.

Fishbach, Ayelet, Janina Steinmetz, and Yanping Tu. 2016. "Motivation in a Social Context: Coordinating Personal and Shared Goal Pursuits with Others." *Advances in Motivation Science* 3: 35–79.

Glewwe, Paul, and Karthik Muralidharan. (2016). "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education* Vol. 5, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 653–743. North-Holland: Elsevier.

Goldberg, Lewis R. 1990. "An Alternative 'Description of Personality': The Big-Five Factor Structure." *Journal of Personality and Social Psychology* 59(6): 1216–1229.

_____. 1992. "The Development of Markers for the Big-Five Factor Structure." *Psychological Assessment* 4(1): 26–42.

Gong, Jie, Yi Lu, and Hong Song. 2021. "Gender Peer Effects on Students' Academic and Noncognitive Outcomes: Evidence and Mechanisms." *Journal of Human Resources* 56(3): 686–710.

Hanushek, Eric A. 2003. "The Failure of Input-based Schooling Policies." *Economic Journal* 113(485): F64–F98.

Hänze, Martin, and Roland Berger. 2007. "Cooperative Learning, Motivational Effects, and Student Characteristics: An Experimental Study Comparing Cooperative Learning and Direct Instruction in 12th Grade Physics Classes." *Learning and Instruction* 17(1): 29–41.

Hart, Jason W., Mark F. Stasson, John M. Mahoney, and Paul Story. 2007. "The Big Five and Achievement Motivation: Exploring the Relationship between Personality and a Two-Factor Model of Motivation." *Individual Differences Research* 5(4): 267–274.

Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annual Review of Economics* 6(1): 689–733.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91(2): 145–149.

Hess, Simon. 2017. "Randomization Inference with Stata: A Guide and Software." *Stata Journal* 17(3): 630–651

Jain, Tarun, and Mudit Kapoor. 2015. "The Impact of Study Groups and Roommates on Academic Performance." *Review of Economics and Statistics* 97(1): 44–54.

Johnson, David W., and Roger T. Johnson. 2002. "Learning Together and Alone: Overview and Meta-analysis." *Asia Pacific Journal of Education* 22(1): 95–105.

Judge, Timothy A., and Remus Ilies. 2002. "Relationship of Personality to Performance Motivation: A Meta-analytic Review." *Journal of Applied Psychology* 87(4): 797–807.

Lavy, Victor, and Edith Sand. 2019. "The Effect of Social Networks on Students' Academic and Noncognitive Behavioural Outcomes: Evidence from Conditional Random Assignment of Friends in School." *Economic Journal*

129(617): 439–480.

Li, Li, Eric Mak, and Chunchao Wang. 2019. “Deskmates: Reconsidering Optimal Peer Assignments within the Classroom.” SSRN 3537509.

Li, Tao, Li Han, Linxiu Zhang, and Scott Rozelle. 2014. “Encouraging Classroom Peer Interactions: Evidence from Chinese Migrant Schools.” *Journal of Public Economics* 111: 29–45.

Lu, Fangwen, and Michael L. Anderson. 2015. “Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups.” *Journal of Labor Economics* 33(1): 91–122.

Malecki, Christine Kerres, and Stephen N. Elliot. 2002. “Children’s Social Behaviors as Predictors of Academic Achievement: A Longitudinal Analysis.” *School Psychology Quarterly* 17(1): 1–23.

Muralidharan, Karthik. 2017. “Field Experiments in Education in Developing Countries.” In *Handbook of Economic Field Experiments* Vol. 2, ed. Abhijit Vinayak Banerjee and Esther Duflo, 323–385. North-Holland: Elsevier.

Neidell, Matthew, and Jane Waldfogel. 2010. “Cognitive and Noncognitive Peer Effects in Early Education.” *Review of Economics and Statistics* 92(3): 562–576.

Roberts, Brent W., Jing Luo, Daniel A. Briley, Philip I. Chow, Rong Su, and Patrick L. Hill. 2017. “A Systematic Review of Personality Trait Change through Intervention.” *Psychological Bulletin* 143(2): 117–141.

Sacerdote, Bruce. 2011. “Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?” In *Handbook of the*

Economics of Education Vol. 3, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 249–277. North-Holland: Elsevier.

Slavin, Robert E. 1980. "Cooperative Learning." *Review of Educational Research* 50(2): 315–342.

Vogl, Tom S. 2014. "Height, Skills, and Labor Market Outcomes in Mexico." *Journal of Development Economics* 107: 84–96.

Wentzel, Kathryn R. 1991. "Relations between Social Competence and Academic Achievement in Early Adolescence." *Child Development* 62(5): 1066–1078.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank.

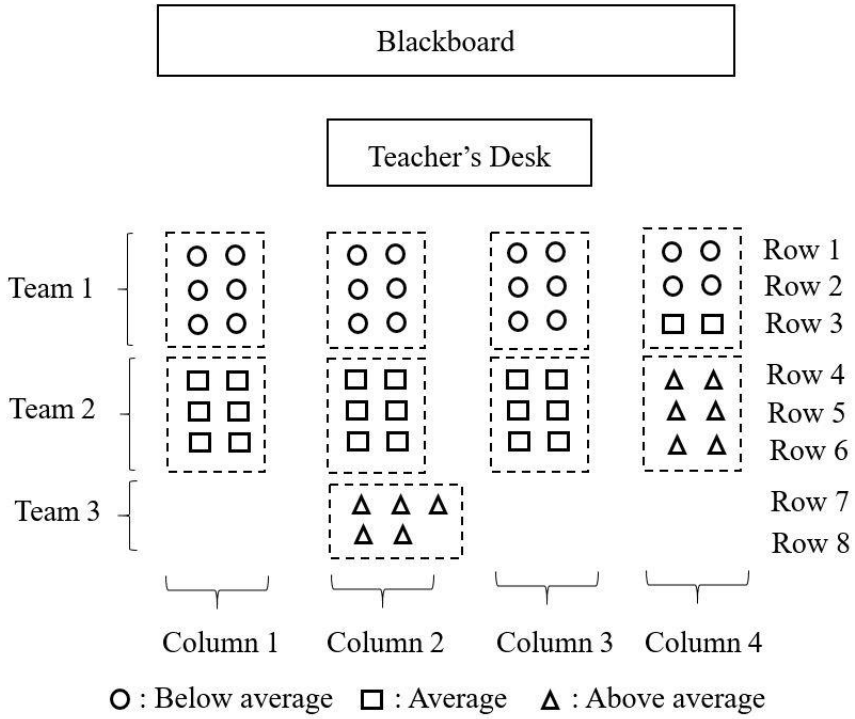

Figure 1
Seating Arrangement in a Typical Classroom

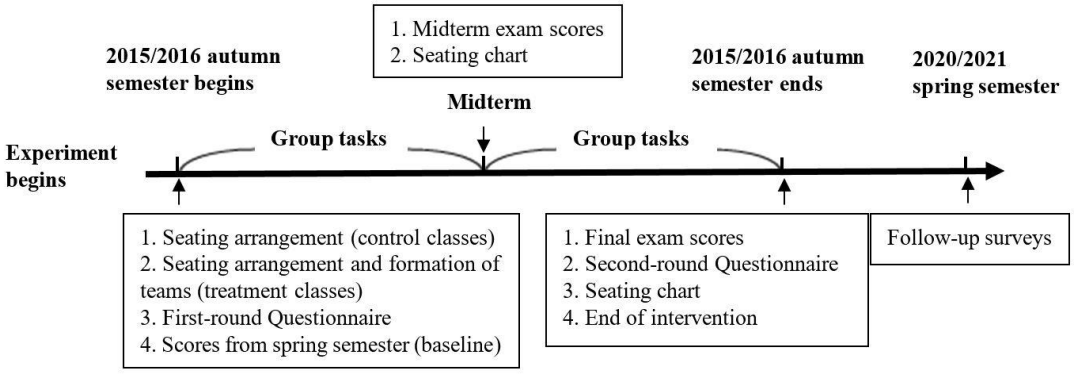

Figure 2
Experimental Procedure

Table 1

Gender Composition of Sample Schools

School code	Total no. of students	Number		Percentage		Treatment class code	Control class code
		Male	Female	Male	Female		
HN 01	258	136	122	53%	47%	III (2) IV (2) V (2)	III (1) IV (1) V (1)
HN02	304	167	137	55%	45%	III (2) IV (2) V (2)	III (1) IV (1) V (1)
HN04	338	187	151	55%	45%	III (2) IV (2) V (1)	III (4) IV (1) V (2)
HN05	365	218	147	60%	40%	III (2) IV (2) V (2)	III (1) IV (1) V (1)
HN06	324	199	125	61%	39%	III (117) IV (113) V (111)	III (118) IV (115) V (112)
Total	1589	907	682	57%	43%		

Data source: Experimental data.

Note: In last two columns, Roman numerals denote grade years, and Arabic numbers enclosed in parentheses refer to class codes.

Table 2

Descriptive Statistics of Treatment and Control Classes

Variables	Treatment classes			Control Classes			Mean Differences
	Mean	Standard deviation	N	Mean	Standard deviation	N	
Average baseline score	55.112	18.418	790	55.260	19.308	799	-0.148
Noncognitive skills							
Openness	0.015	1.018	775	-0.015	0.982	783	0.030
Conscientiousness	0.011	1.026	775	-0.012	0.977	783	0.023
Extraversion	-0.011	1.018	775	0.013	0.985	783	-0.024
Agreeableness	-0.049	1.040	775	0.049	0.957	788	-0.098*
Neuroticism	0.074	1.011	775	-0.077	0.035	783	0.151***
Individual characteristics							
Gender (1=male, 0=female)	0.568	0.496	790	0.573	0.495	799	-0.005
Age (year)	9.501	1.070	790	9.503	1.009	799	-0.002
Height (cm)	135.306	11.621	790	134.585	9.463	799	0.722

Data source: Experimental data.

Note: Noncognitive skills are standardized to have a mean of 0 and a s.d. of 1. *** and * indicate significance at 1% and 10%, respectively.

Table 3

Descriptive Statistics of Variables

Variables	Mean	Standard deviation	Min	Max	N
Cognitive skills					
Standardized baseline scores	0	1	-2.93	2.08	1589
Standardized midterm scores	0	1	-3.02	2.33	1589
Standardized final scores	0	1	-3.01	2.27	1589
Noncognitive skills					
Openness	0	1	-4.95	5.26	3043
Conscientiousness	0	1	-3.61	4.63	3043
Extraversion	0	1	-4.03	5.27	3043
Agreeableness	0	1	-2.66	4.29	3043
Neuroticism	0	1	-4.9	3.36	3043
Control variables					
Treatment × Post	0.25	0.43	0	1	3178
Treatment team dummy	0.50	0.50	0	1	3178
Extracurricular reading time	11.74	17.56	0	300	3036
Household chore time	12.36	20.51	0	288	3035
Parents' yearly income	1.89	0.59	0	5.19	3011

Data source: Experimental data.

Note: Extracurricular reading time and household chore time are measured in minutes per day. Parents' yearly income is the logarithm of RMB 10,000.

Table 4

Effects of Forming Teams on Cognitive Skills

	Midterm Examination				Final Examination			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS Estimate								
Treatment	0.073***	0.077***	0.078***	0.095***	0.072***	0.075***	0.076***	0.082***
	(0.013)	(0.013)	(0.014)	(0.018)	(0.018)	(0.020)	(0.021)	(0.025)
Baseline		0.816***	0.816***	0.810***		0.791***	0.791***	0.788***
		(0.026)	(0.026)	(0.027)		(0.024)	(0.024)	(0.024)
Individual characteristics		yes	yes	yes		yes	yes	yes
Class characteristics			yes	yes			yes	yes
Family characteristics				yes				
N	1589	1589	1589	1462	1589	1589	1589	1462
Adjusted R^2	0.004	0.700	0.699	0.695	0.001	0.685	0.683	0.686
Panel B: DID Estimate								
Treatment	-0.018	-0.012	-0.011	-0.011	-0.017	-0.012	-0.009	-0.009
	(0.014)	(0.012)	(0.012)	(0.014)	(0.014)	(0.012)	(0.011)	(0.012)
Post	-0.023*	-0.023*	-0.024	-0.026	-0.033	-0.033	-0.032	-0.028
	(0.014)	(0.012)	(0.015)	(0.015)	(0.020)	(0.020)	(0.021)	(0.021)
Treatment	0.090***	0.090***	0.091***	0.100***	0.089***	0.089***	0.087***	0.085***
× Post	(0.015)	(0.015)	(0.015)	(0.018)	(0.023)	(0.023)	(0.023)	(0.025)
Baseline		0.909***	0.909***	0.908***		0.897***	0.897***	0.898***
		(0.012)	(0.012)	(0.013)		(0.012)	(0.0123)	(0.012)
Individual characteristics		yes	yes	yes		yes	yes	yes
Class characteristics			yes	yes			yes	yes
Family characteristics				yes				
N	3178	3178	3178	3003	3178	3178	3178	3003
Adjusted R^2	0.001	0.829	0.828	0.828	0.001	0.82	0.82	0.823
Finite sample inference p				0.000				0.000

Note: Each column is a separate regression. Panel A reports the OLS estimates using test scores in midterm examination (columns (1) – (4)) and final examination (columns (5) – (8)) as dependent variables. Panel B reports the results of DID estimates from Equation (1) using midterm examination scores in columns (1) – (4) and final examination scores in columns (5) – (8). Individual characteristics are gender, age, and height. Class characteristics are school, grade, class, and team dummy variables. Family characteristics are parents' income, extracurricular reading time, and household chore time. Standard errors are clustered at the school-grade level and enclosed in parentheses. Finite sample inference is reported in the last row. *** and * indicate significance at 1% and 10%, respectively.

Table 5

Heterogeneous Effects of Forming Teams: Cognitive Skills

	Gender		Academic Performance				Grade		
	Male	Female	A	B	C	D	Third	Fourth	Fifth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Midterm	0.128***	0.065*	0.240	0.046	0.043	0.086**	0.186***	0.116*	0.037**
	(0.035)	(0.035)	(0.184)	(0.052)	(0.364)	(0.041)	(0.019)	(0.038)	(0.011)
Between group test	p(female=male)=0.3001		p(A=B)=0.4957 p(A=C)=0.5727 p(A=D)=0.5377 p(B=C)=0.9694 p(B=D)=0.4338 p(C=D)=0.4448				p(third=fourth)=0.3666 p(third=fifth)=0.0322 p(fourth=fifth)=0.3107		
Final	0.057	0.105*	0.276	0.068	0.047	0.063**	0.165***	0.119	0.003
	(0.043)	(0.049)	(0.236)	(0.082)	(0.068)	(0.031)	(0.024)	(0.059)	(0.020)
Between group test	p(female=male)=0.4294		p(A=B)=0.8110 p(A=C)=0.9432 p(A=D)=0.3412 p(B=C)=0.7995 p(B=D)=0.3135 p(C=D)=0.3654				p(third=fourth)=0.5594 p(third=fifth)=0.0221 p(fourth=fifth)=0.1322		

Note: Each column reports the results of a cohort with the specification including individual characteristics, school characteristics, and family characteristics controlled. Midterm represents the DID analysis comparing midterm examination scores with baseline scores. Final represents the DID analysis comparing final examination scores with baseline scores. A, B, C, and D denote levels of academic performance from high to low. Standard errors are clustered at the school-grade level and enclosed in parentheses. Between-group test reports the p-value from the between-group t-test. ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Table 6

Between-group Comparison of Noncognitive Skills

	First-round questionnaire			Second-round questionnaire		
	Treatment classes	Control classes	Mean differences	Treatment classes	Control classes	Mean differences
Openness	0.015	-0.015	0.030	0.034	-0.035	0.070
Conscientiousness	0.011	-0.011	0.023	0.112	-0.107	0.219***
Extraversion	-0.011	0.013	-0.024	0.066	-0.063	0.129**
Agreeableness	-0.049	0.049	-0.098*	-0.073	0.068	-0.140***
Neuroticism	0.074	-0.077	0.151***	0.115	-0.108	0.223***

Data source: Experimental data.

Note: ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Table 7

Effects of Forming Teams on Noncognitive Skills

	OLS estimates				DID estimates				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Agreeableness	-0.138*	-0.127*	-0.147***	-0.141**	-0.116*	-0.122*	-0.133**	-0.149**	0.020
	(0.075)	(0.073)	(0.055)	(0.054)	(0.063)	(0.063)	(0.062)	(0.063)	
Conscientiousness	0.219***	0.215**	0.222**	0.217**	0.209**	0.207***	0.154**	0.140*	0.000
	(0.052)	(0.105)	(0.105)	(0.098)	(0.087)	(0.072)	(0.070)	(0.078)	
Extraversion	0.129**	0.118	0.123	0.115	0.156	0.147	0.140	0.130	0.100
	(0.052)	(0.099)	(0.111)	(0.122)	(0.109)	(0.122)	(0.116)	(0.109)	
Neuroticism	0.223***	0.214***	0.229***	0.226***	0.170***	0.172***	0.162**	0.200***	0.010
	(0.076)	(0.075)	(0.062)	(0.063)	(0.065)	(0.065)	(0.065)	(0.065)	
Openness	0.070	0.072	0.060	0.065	0.041	0.046	0.056	0.057	0.190
	(0.052)	(0.089)	(0.087)	(0.091)	(0.030)	(0.036)	(0.041)	(0.035)	
Individual Characteristics		yes	yes	yes		yes	yes	yes	
School Characteristics			yes	yes			yes	yes	
Family Characteristics				yes				yes	

Note: Each column is a separate regression. Columns (1) – (4) use OLS specifications using only the personality traits from the second-round questionnaire, and columns (5) – (8) use DID specifications. Column (9) reports the finite sample inference p-value of the full specification in column (8). Individual characteristics are gender, age, and height. Class characteristics are school, grade, class, and team dummy variables. Family characteristics are parents' income, extracurricular reading time, and household chore time. Standard errors are clustered at the school-grade level and enclosed in parentheses. ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Table 8

Heterogeneous Effects of Forming Teams: Noncognitive Skills

	Gender		Academic Performance				Grade		
	Male	Female	A	B	C	D	Third	Fourth	Fifth
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Openness	0.105	0.003	0.092	-0.045	0.149	-0.315	0.104	0.038	0.125
	(0.146)	(0.125)	(0.101)	(0.215)	(0.291)	(0.500)	(0.150)	(0.167)	(0.170)
Between group test	p(female=male)=0.1306		p(A=B)=0.3093 p(A=C)=0.9650 p(A=D)=0.7003 p(B=C)=0.5421 p(B=D)=0.8993 p(C=D)=0.7635				p(third=fourth)=0.9963 p(third=fifth)=0.4476 p(fourth=fifth)=0.4671		
conscientiousness	0.174	0.116	-0.266	-0.134	0.179	0.199**	0.297**	0.303*	0.023
	(0.112)	(0.131)	(0.825)	(0.271)	(0.158)	(0.088)	(0.145)	(0.176)	(0.151)
Between group test	p(female=male)=0.7086		p(A=B)=0.4304 p(A=C)=0.1079 p(A=D)=0.2497 p(B=C)=0.3931 p(B=D)=0.4318 p(C=D)=0.7902				p(third=fourth)=0.6363 p(third=fifth)=0.4461 p(fourth=fifth)=0.2268		
Extraversion	0.208*	0.050	0.303	0.195	0.029	0.149	0.297**	0.303*	0.023
	(0.117)	(0.129)	(0.831)	(0.243)	(0.158)	(0.126)	(0.145)	(0.176)	(0.151)
Between group test	p(female=male)=0.3260		p(A=B)=0.5783 p(A=C)=0.7644 p(A=D)=0.2254 p(B=C)=0.5289 p(B=D)=0.1698 p(C=D)=0.3382				p(third=fourth)=0.9613 p(third=fifth)=0.4389 p(fourth=fifth)=0.5097		
Agreeableness	-0.092	-0.183*	-0.156	-0.229	-0.222*	-0.113	0.009	-0.022	-0.271**
	(0.088)	(0.101)	(0.779)	(0.224)	(0.125)	(0.079)	(0.173)	(0.151)	(0.127)
Between group test	p(female=male)=0.2323		p(A=B)=0.3513 p(A=C)=0.3222 p(A=D)=0.9228 p(B=C)=0.8259 p(B=D)=0.6562 p(C=D)=0.5861				p(third=fourth)=0.6427 p(third=fifth)=0.0923 p(fourth=fifth)=0.0454		
Neuroticism	0.122	0.231**	-0.095	0.396*	0.201	0.151*	0.169	0.092	0.224*
	(0.090)	(0.105)	(0.600)	(0.218)	(0.138)	(0.078)	(0.119)	(0.209)	(0.125)
Between group test	p(female=male)=0.2596		p(A=B)=0.2765 p(A=C)=0.2739 p(A=D)=0.6667 p(B=C)=0.8286 p(B=D)=0.4035 p(C=D)=0.3590				p(third=fourth)=0.1142 p(third=fifth)=0.6508 p(fourth=fifth)=0.0584		

Note: Each column reports the results of a cohort with the specification including individual characteristics, school characteristics, and family characteristics controlled. A, B, C, and D denote levels of academic performance from high to low. Standard errors are clustered at the school-grade level and enclosed in parentheses. The between-group test reports the p-value from the between-group t-test. ** and * indicate significance at 5% and 10%, respectively.

Table 9

Between-Group Comparison of Class Behaviors

	First round			Second round		
In-class discipline and attention	Treatment classes	Control classes	Mean differences	Treatment classes	Control classes	Mean differences
Unpermitted talking	3.225	3.867	-0.642**	3.392	6.857	-3.465***
Pranking	3.094	3.909	-0.815***	3.165	6.304	-3.139***
Reading comics	0.741	0.170	0.571***	0.997	1.556	-0.558*
Sleeping	0.250	0.088	0.162***	1.125	1.472	-0.347
Tardiness	0.106	0.062	0.043**	1.108	1.497	-0.388
In-class attention	1.547	1.489	0.059	1.693	1.503	0.190***

Note: The measurement for talking without permission, pranking, reading comic books in class, sleeping in class, and tardiness is frequency per week. In-class attention is measured on three levels, that is, 0 = mostly cannot pay attention, 1 = not sure, or 2 = mostly can pay attention. ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Table 10

Effects of Forming Teams on In-class Discipline and Class Attention

	Talk	Prank	Comics	Sleep	Tardiness	Attention
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment× Post	-1.961***	-1.563***	-0.801***	0.000	-0.006	0.234***
	(0.293)	(0.287)	(0.303)	(0.013)	(0.016)	(0.047)
Treatment	0.218	-0.172	0.274**	0.016*	0.028**	-0.021
	(0.228)	(0.234)	(0.130)	(0.009)	(0.011)	(0.039)
Post	1.189***	0.748***	1.106***	0.018**	0.024***	-0.023
	(0.200)	(0.194)	(0.212)	(0.008)	(0.009)	(0.035)
Individual characteristics	yes	yes	yes	yes	yes	yes
Class characteristics	yes	yes	yes	yes	yes	yes
Family characteristics	yes	yes	yes	yes	yes	yes
N	2580	2580	2669	2469	2489	2049
Adjusted R^2	0.153	0.170	0.053	0.017	0.021	0.200

Note: Each column reports the results of a separate equation. The measurement for talking without permission, pranking, reading comic books in class, sleeping in class, and tardiness is times per week. In-class attention is measured on three levels: 0 = mostly cannot pay attention, 1 = not sure, and 2 = mostly can pay attention. Individual characteristics are gender, age, and height. Class characteristics are school, grade, class, and team dummy variables. Family characteristics are parents' income, extracurricular reading time, and household chore time. Standard errors are clustered at the school-grade level and enclosed in parentheses. ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Table 11

Descriptive Statistics of Treatment Classes and Survey Sample

Variables	Treatment classes			Random sample			Mean differences
	Mean	Standard deviation	N	Mean	Standard deviation	N	
Cognitive skills							
Standardized baseline scores	-0.001	0.991	791	-0.004	1.005	73	0.003
Non-cognitive skills							
Openness	0.015	1.019	728	-0.031	0.881	73	0.065
Conscientiousness	0.112	1.054	728	0.062	0.795	73	0.051
Extraversion	0.067	1.074	728	0.085	0.957	73	-0.018
Agreeableness	-0.074	1.087	728	-0.174	0.933	73	0.100
Neuroticism	0.116	1.047	728	0.094	0.868	73	0.022
Individual characteristics							
Gender	0.568	0.496	790	0.589	0.495	73	-0.021
Age	9.501	1.070	790	9.521	1.042	73	-0.020
Height	135.306	11.621	790	135.658	8.39	73	-0.352
Grade	4.006	0.821	790	3.986	0.842	73	0.020
School	3.061	1.448	790	3.123	1.443	73	-0.063

Data source: Experiment data.

Note: Cognitive skills and non-cognitive skills are standardized to have a mean of 0 and a s.d of 1.

Table 12

Follow-up Survey on Class Teachers

Question	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total
Connection	9 (60%)	5 (33%)	1 (7%)	0 (0%)	0 (0%)	15
Mutual help	6 (40%)	8 (53%)	1 (7%)	0 (0%)	0 (0%)	15
Study supervision	8 (53%)	5 (33%)	2 (13%)	0 (0%)	0 (0%)	15
Discipline supervision	3 (20%)	10 (67%)	2 (13%)	0 (0%)	0 (0%)	15
Competition	6 (40%)	8 (53%)	1 (7%)	0 (0%)	0 (0%)	15
Test score	7 (47%)	5 (33%)	2 (13%)	1 (7%)	0 (0%)	15
In-class discipline	2 (13%)	8 (53%)	4 (27%)	1 (7%)	0 (0%)	15
Attention	0 (0%)	2 (13%)	5 (33%)	8 (53%)	0 (0%)	15
Curriculum	0 (0%)	1 (7%)	4 (27%)	10 (67%)	0 (0%)	15
Pedagogy	0 (0%)	0 (0%)	5 (33%)	10 (67%)	0 (0%)	15
Other class	0 (0%)	0 (0%)	7 (47%)	8 (53%)	0 (0%)	15

Data source: Survey data.

Note: For each entry, the number is the count of the corresponding answer. Percentage in parenthesis is the count of corresponding answer divided by total count of all answers.

Table 13

Follow-up Survey on Students

Question	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total
Adapt time	22 (30%)	36 (49%)	13 (18%)	2 (3%)	0 (0%)	73
Make friends	22 (30%)	40 (55%)	7 (10%)	4 (5%)	0 (0%)	73
Team continue	19 (26%)	36 (49%)	15 (21%)	3 (4%)	0 (0%)	73
Responsible	36 (49%)	32 (44%)	4 (5%)	1 (1%)	0 (0%)	73
Think action	28 (38%)	40 (55%)	5 (7%)	0 (0%)	0 (0%)	73
More effort	33 (45%)	33 (45%)	7 (10%)	0 (0%)	0 (0%)	73
Friend easily	31 (42%)	33 (45%)	9 (12%)	0 (0%)	0 (0%)	73
After school	28 (38%)	31 (42%)	12 (16%)	1 (1%)	1 (1%)	73
Study self-policing	32 (44%)	34 (47%)	7 (10%)	0 (0%)	0 (0%)	73
Discipline self-policing	35 (48%)	33 (45%)	4 (5%)	1 (1%)	0 (0%)	73
Study from stigma	23 (32%)	32 (44%)	15 (21%)	3 (4%)	0 (0%)	73
Discipline from stigma	28 (38%)	28 (38%)	15 (21%)	2 (3%)	0 (0%)	73
Study role model	33 (45%)	29 (40%)	10 (14%)	1 (1%)	0 (0%)	73
Discipline role model	33 (45%)	30 (41%)	9 (12%)	1 (1%)	0 (0%)	73
Cooperation	35 (48%)	32 (44%)	6 (8%)	0 (0%)	0 (0%)	73
Kolmogorov-Smirnov test on reasons for studying						
p(self policing=social stigma)=0.379						
p(self policing=role model)=1.000						
p(social stigma=role model)=0.500						
Kolmogorov-Smirnov test on reasons for observing in-class discipline						
p(self policing=social stigma)=0.277						
p(self policing=role model)=0.996						
p(social stigma=role model)=0.890						

Data source: Survey data.

Note: For each entry, the number is the count of the corresponding answer. Percentage in parentheses is the count of the corresponding answer divided by the total count of all answers.

Table 14

Correlation between Baseline Test Scores and Noncognitive Skills

	(1)	(2)	(3)	(4)	(5)	(6)
Openness	0.118***					0.048**
	(0.018)					(0.017)
Conscientiousness		0.211***				0.156***
		(0.030)				(0.034)
Extraversion			0.136***			0.037
			(0.026)			(0.032)
Agreeableness				0.109**		0.039
				(0.039)		(0.031)
Neuroticism					-0.143***	-0.082***
					(0.024)	(0.021)
Individual characteristics	yes	yes	yes	yes	yes	yes
Class characteristics	yes	yes	yes	yes	yes	yes
Family characteristics	yes	yes	yes	yes	yes	yes
N	1540	1540	1540	1540	1540	1540
Adjusted R^2	0.055	0.083	0.059	0.051	0.059	0.094

Note: Each column is a separate regression. Individual characteristics are gender, age, and height. Class characteristics are school, grade, class, and team dummy variables. Family characteristics are parents' income, extracurricular reading time, and household chore time. Standard errors are clustered at the school-grade level and enclosed in parentheses. *** and ** indicate significance at 1% and 5%, respectively.

Table 15

Treatment Effect with Controls on Noncognitive Skills

	(1)	(2)	(3)	(4)
Treatment	-0.009	-0.001	0.001	0.000
	(0.036)	(0.003)	(0.004)	(0.006)
Post	-0.020	0.000	0.001	-0.001
	(0.037)	(0.005)	(0.005)	(0.004)
Treatment× Post	0.061*	0.079***	0.080***	0.083***
	(0.034)	(0.018)	(0.019)	(0.019)
Openness	0.068***	0.012*	0.015**	0.016**
	(0.018)	(0.006)	(0.006)	(0.006)
Conscientiousness	0.116***	0.005	0.004	0.004
	(0.013)	(0.007)	(0.008)	(0.008)
Extraversion	0.026	-0.002	-0.002	-0.001
	(0.023)	(0.007)	(0.007)	(0.007)
Agreeableness	-0.004	0.001	0.006	0.005
	(0.025)	(0.007)	(0.007)	(0.007)
Neuroticism	-0.075***	0.001	0.000	0.000
	(0.025)	(0.008)	(0.009)	(0.009)
Individual characteristics		yes	yes	yes
Class characteristics				yes
Family characteristics				yes
N	3045	2988	2988	2980
Adjusted R^2	0.039	0.850	0.850	0.850

Note: Each column is a separate regression. Individual characteristics are gender, age, and height. Class characteristics are school, grade, class, and team dummy variables. Family characteristics are parents' income, extracurricular reading time, and household chore time. Standard errors are clustered at the school-grade level and enclosed in parentheses. ***, **, and * indicate significance at 1%, 5%, and 10%, respectively.

Appendix: Follow-up Survey

We perform a Runs test on the selected sample. The results are presented in Table A1. No evidence of violation of the randomness was observed.

Table A1

Runs Test on Survey Sample

Variables	Test Value	Cases< test value	Cases≥ test value	Total cases	Number of runs	Z score	p value
Baseline scores	-0.004	33	38	71	30	-1.520	0.129
Noncognitive skills							
Openness	-0.031	37	35	72	42	1.194	0.232
Conscientiousness	0.062	32	40	72	42	1.309	0.191
Extraversion	0.085	36	36	72	36	-0.237	0.812
Agreeableness	-0.174	31	41	72	35	-0.316	0.752
Neuroticism	0.094	34	38	72	34	-0.688	0.492
Individual characteristics							
Gender	0.589	30	43	73	29	-1.788	0.074
Age	9.521	36	37	73	31	-1.531	0.126
Height	137.658	36	37	73	38	0.119	0.905
Grade	3.986	26	47	73	28	-1.667	0.095
School	3.123	43	30	73	30	-1.545	0.122

Data source: Experiment data.

Note: Baseline scores and non-cognitive skills are standardized to have a mean of 0 and a s.d of 1. Gender is a dummy variable which is 1 for male and 0 for female. Height is measured in centimeters. School is a category variable with values 1 to 5.

In the survey on the class teachers, we asked them seven questions about their observations on the changes in the students' behaviors. In addition, four questions were about the general experiment implementation. The questions are as follow:

1. I observed a close connection between the student team members.

2. I observed mutual help for studying among the student team members.

3. I observed mutual supervision on studying among the student team members.

4. I observed mutual supervision on in-class discipline among the student team members.

5. I observed strengthened awareness of competition among the students.

6. The student teams were helpful in improving academic performance.

7. The student teams were helpful in maintaining in-class discipline.

8. I paid more attention to the treatment classes than to the control classes during the experiment.

9. I made a separate curriculum for treatment classes.

10. I adopted a different teaching pedagogy for the treatment classes.

11. I organized team activities in the control classes during the experiment.

In the survey on the students, we instructed them to reflect on their behavioral changes during the experiment. The questions are as follow:

1. After being assigned to a team, it took me some time to adjust to the team activities.

2. After being assigned to a team, it took me some time to make friends with the other team members.

3. If possible, I would like to have the team activities continue after the experiment.

4. After being assigned to a team, I believed that I should be responsible for the team appraisals.

5. After being assigned to a team, I considered whether my actions would affect the team appraisals.

6. After being assigned to a team, I exerted efforts to receive positive team appraisals.

7. The student teams provided me with an environment to make friends easily.

8. I liked interacting with the team members even after school.

9. After being assigned to a team, I studied hard voluntarily.

10. After being assigned to a team, I observed in-class discipline voluntarily.

11. After being assigned to a team, I studied hard, because the team members would put pressure on me if I did not do so.

12. After being assigned to a team, I observed in-class discipline, because the team members would put pressure on me if I did not do so.

13. After being assigned to a team, I studied hard, because one of the team members was my role model.

14. After being assigned to a team, I observed in-class discipline, because one of the team members was my role model.

15. My awareness of cooperation increased after I joined the team.