I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15944

The Impact of Computer-Assisted Instruction on Student Performance: Evidence from the Dual-Teacher Program

Haizheng Li
Zhiqiang Liu
Fanzheng Yang
Li Yu

FEBRUARY 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15944

The Impact of Computer-Assisted Instruction on Student Performance: Evidence from the Dual-Teacher Program

Haizheng Li
Georgia Institute of Technology and IZA

Zhiqiang Liu
University at Buffalo

Fanzheng Yang
Central University of Finance and Economics, Beijing

Li Yu
Central University of Finance and Economics, Beijing

FEBRUARY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Impact of Computer-Assisted Instruction on Student Performance: Evidence from the Dual-Teacher Program*

We present findings from an evaluation study of the Dual-Teacher program, a computer-assisted instruction program, that makes lecture videos and other teaching resources from an elite urban middle school available through the internet to schools in poor and remote areas in China. The unique design of the study allows us to not just estimate the effect of the program on student performance but distinguish the direct effect coming from students' exposure to the lecture videos in class and the indirect effect due to improved instruction quality of the local teacher who uses the lecture videos in lesson preparation. Using the difference-in-differences method, we find that the Dual-Teacher program improves student performance in math by 0.978 standard deviations over the three-year middle school education, of which 0.343 standard deviations are attributable to the indirect effect. We also find that the positive impacts of the program are cumulative and robust to student and teacher characteristics as well as a plethora of other considerations. From a policy perspective, our findings suggest that the Dual-Teacher program is an effective and low cost means to improve education outcomes in underserved areas and hence help close cross-region gaps in education.

JEL Classification: I21, I24, I28, O15

Keywords: computer-assisted instruction, computer-assisted learning, education policy, inequality

Corresponding author:
Zhiqiang Liu
Department of Economics
University at Buffalo
Buffalo, New York
USA
E-mail: zqliu@buffalo.edu

* Partial financial support was provided by the National Natural Science Foundation of China (Grant #71773151). We thank Min Tang, Qing Miao and Ming Wen at the YouChange China Social Entrepreneur Foundation for their support and inputs and the Dual-Teacher Program participating schools for their assistance in data collection. Thanks also go to Belton Fleisher, Jiapeng Dong, Ning Jia, Jinlan Ni, Bing Wang, Xuefei Wang, Nina Yin and participants at the 92nd Annual Meeting of the Southern Economic Association, November 2022, Fort Lauderdale, Florida for their comments and suggestions. We would like to express our gratitude to Xin Chen and Yuzhe Ning, two former Ph.D. students at the China Center for Human Capital and Labor Market Research, for excellent assistance in survey design, field visit, and data collection and processing. We are responsible for remaining errors.

1. Introduction

Large cross-region gaps in education are very common around the world, especially in developing countries, such as China (Fleisher et al., 2010), India (Desai and Kulkarni, 2008), Mexico (Navarro-Sola, 2021), and African countries (Zhang, 2006). Numerous programs designed to reduce the inequality, such as free textbooks (Glewwe et al., 2009), teacher training (Harris and Sass, 2011; Loyalka et al., 2019), subsidies to high-quality teachers for relocation to rural areas (Fowler, 2003; Prince, 2002), remedial education (Banerjee et al. 2007; Jacob and Lefgren, 2004; Lavy and Schlosser, 2005; Battaglia and Lebedinski, 2015), computer-assisted learning (Banerjee et al. 2007; Bettinger et al., 2020; Lai et al., 2015; Ma et al., 2020; Mo et al., 2014; Mo et al., 2020), and computer-assisted instruction (Beg et al., 2019; Bianchi et al., 2022), [2] have been shown to be effective to varying degrees, ranging from no positive effect to highly significant impact on learning outcomes of students in underserved areas. The advent and widespread application of the internet technology have created new methods and opportunities to close cross-region gaps in education. Online education and distance teaching make high-quality instructional services accessible at lower costs than traditional modes that require the physical presence of high-quality teachers in the classroom, and thus provide a promising solution to inadequacy of educational resources in poor and remote areas (Naik et al., 2020; Wennersten et al., 2015).

However, distance education cannot completely replace the traditional face-to-face education due to some inherent limitations of the former. First, compared with face-to-face instruction, online lectures lack live teacher-student interactions, such as classroom discussions, led by the teacher and therefore may undermine the effectiveness of student learning (Alpert et al., 2016; Figlio et al., 2013). Second, without real-time feedback from students, the teacher in remote location is unable to adjust the content as well as the pace of their lectures according to the level of comprehension on the part of students.[3] Third, online lectures are often tailored

[2] Computer-assisted learning (CAL) is different from computer-assisted instruction (CAI), although some authors use the terms interchangeably. CAL focuses more on a direct link connecting students to learning through technology like apps or educational software that students use outside regular classes to reinforce and supplement what they learned in class, whereas CAI provides computer-based instructional assistance to teachers to enhance or complement their in-class instructions and contents.

[3] Johnston and Ksoll (2017), and Naik et al. (2020) evaluate the impact of remote instruction via satellite in Ghana and India, respectively. Both studies discuss a particular type of remote teaching that allows students to communicate directly and in real time with a live instructor, whereas other online interventions typically use static instructional aids.

toward students of certain background and therefore may not be suitable for students with diverse range of abilities in underserved areas. It has been shown that the pedagogy of "teaching at the right level" is a key to good learning outcomes (Banerjee et al., 2010; Banerjee et al., 2016).

To address the deficiencies of pure online teaching, policy makers and educational experts have recently begun to promote blended teaching programs that combine modern online teaching and traditional offline instruction by partially replacing or supplementing in-person instruction with remote lectures, such as live broadcasted lessons, pre-recorded lecture videos, and TV shows. Several studies (Beg et al., 2019; Bianchi et al., 2022; Borzekowski, 2018; Borzekowski and Henry, 2011; Borzekowski et al., 2019; Näslund-Hadley et al., 2014; Wennersten et al., 2015) have provided evidence suggesting a positive role of hybrid teaching programs in improving educational outcomes in underserved regions where local teachers may not fully master the subject matter they are expected to teach or may not have sufficient teaching skills to deliver effective lectures.

The aim of this study is to evaluate a novel hybrid program, called the Dual-Teacher program (DT program hereafter) that leverages the internet technology to make high quality teaching resources accessible to schools in poor and remote areas of China. Specifically, a teacher in a remote area school gains access to teaching resources from an elite middle school in Beijing, including mainly recorded lecture videos, and incorporates them into their own teaching activities in two ways: watching the recorded lecture videos while preparing for their own lectures and showing selectively the recorded lecture videos in class. Therefore students are taught by two teachers.[4] Differing from the practice that brings high-quality lectures live to classrooms in remote schools via the internet to replace in-person lectures, such as Massive Open Online Courses (MOOCs) and Mexico's Telesecondary schools (Borghesan and Vasey, 2021; Navarro-Sola, 2021), in the DT program the local teacher determines how many hours to spend on watching the lecture videos during lesson preparation and how many hours and which parts of the lecture videos to show directly to students in class to supplement their in-person lectures.

[4] The cost of implementing the DT program is relatively small. First, there is little extra cost to remote schools, since the Internet has become widely available even in remote areas and connecting schools entails little extra costs. Second, it does not require costly teacher training while works like a self-administered professional training program for local teachers. Third, there is no extra cost to the elite urban school given that teaching resources in the urban school are already available and the DT program does not entail diverting resources from urban to rural/remote areas.

In this paper, we first develop a theoretical model to illustrate the channels through which the DT program affects student outcomes. The model is based on Acemoglu et al. (2014), in which a "superstar" teacher delivers all the lectures remotely and local teachers provide all other teaching services. We differ from Acemoglu et al. (2014) in that (1) the local teacher also delivers in-person lectures in addition to other services, and (2) the instruction quality of the local teacher improves over time as they study the lecture videos during lesson preparation. In our model, the DT program improves student performance through exposing students to high-quality lecturing by the teacher in the elite urban school and improving instruction quality of the local teacher. Using data that we collected from nine middle schools that participated in the math-focused DT program and a difference-in-differences model, we find that the DT program has a positive and statistically significant effect on student performance in math and the improvement is attributable to high-quality lecturing by the remote teacher as well as improvement in the quality of instruction of the local teacher. Our estimates show that on average the DT program increased math test scores by 0.978 standard deviations over the three-year middle school education, of which 0.343 standard deviations are attributable to improved instruction on the part of the local teacher. We also find that the positive impact of the DT program is cumulative and robust to student and teacher characteristics as well as a plethora of other considerations. Overall, our findings suggest that the DT program is an effective means to improve education outcomes in underserved areas and hence help close cross-region gaps in education.

This study contributes to two strands of the literature. The first is the literature that examines the effects of remote instruction and computer-assisted instruction (CAI) on academic performance of students in underserved schools.[5] A large number of studies, for example Näslund-Hadley et al. (2014), Wennersten et al. (2015), Beg et al. (2019), and Bianchi et al. (2022), find positive impacts of CAI programs on student outcomes. Several evaluation studies of local adaptations of the well-known TV program "Sesame Street" by Borzekowski (2018) in Tanzania, Borzekowski and Henry (2011) in Indonesia, and Borzekowski et al. (2019) in Rwanda also find it to be conducive to developing numeracy and literacy skills of young children. However, Angrist and Lavy (2002) and Goolsbee and Guryan (2006) find that increased use of

[5] Recognizing deficiencies of pure online teaching, many advocate hybrid instruction modes that combine remote and face-to-face lectures (McPherson and Bacow, 2015).

computer and internet in schools do not appear to result in better student performance in Israel and the U.S., respectively.

The current paper is closely related to Bianchi et al. (2022) that investigates the long-term impacts of a large computer-assisted intervention program on cognitive and labor market outcomes of students in rural China. The program that Bianchi et al. study shares several similarities with the DT program: both targeted schools in underserved areas, and both used internet technology to deliver quality instructions and other teaching materials. There are major differences. One is that we have specific information on schools and teachers that participated in the program and students who were exposed to the program in different ways and at different intensities, while Bianchi et al. do not have such information and have to assume a uniform level of compliance across provinces and counties where the program is introduced. Second, Bianchi et al.'s control groups are students of older cohorts and students from a small number of counties in different provinces, while our control groups are students from the same school and cohort. In other words, our treatment effect estimates are obtained from highly homogeneous groups of students, in the absence of the DT program. Third, we estimate the impact of the DT program over one semester to a period of three years, and our results are, therefore, less susceptible to confounding factors, which are more likely to be a concern for analyses of long-term effects of a policy intervention in the difference-in-differences framework. Our study complements Bianchi et al. (2022) in that our findings of short- to medium-term effects help bolster their findings of long-term effects. After all, it is hard to imagine that the positive effects of the program on cognitive outcomes are evident 7 to 10 years later but not during or immediate after the intervention. Furthermore, the unique design of our "quasi-experiment" and detailed information on the characteristics of schools, teachers and students allow us to isolate the effects of direct and indirect exposures to the DT program on student performance and explore variations in treatment intensity. Our paper differs from existing studies about CAL programs (for example, Lai et al., 2015; Ma et al., 2020; Mo et al., 2014; Mo et al., 2020) in two critical aspects. One is that the CAL programs they study are remedial tutoring tools that students use outside the regular class, while ours is an integral part of the in-class instruction. The other difference is in research design. To the best of our knowledge, in the existing CAL studies treatment and control groups are randomly assigned schools, while in our study they are classes from the same school and cohort.

The second strand of the literature that this study is closely related to is on improving teacher quality in underserved areas in general and through professional development in particular. For poor and remote areas in a country, it is relatively easy for the government to improve the hardware of local schools, such as classrooms, the internet and computer facilities. However, it is more difficult to improve the quality of local teachers. Previous studies have investigated various programs designed to increase teacher performance through monetary incentives, such as competitive salary (Loeb and Page, 2000), performance pay (Muralidharan and Sundararaman, 2011) or through certification (Kane et al., 2008), evaluation (Master, 2014; Taylor and Tyler, 2012), and training (Harris and Sass, 2011; Loyalka et al., 2019). These programs may not be effective for schools in poor and remote areas where the pay is low and work environment is undesirable. In this situation, additional incentives such bonuses, stipends, low-interest loans, and housing assistance are used to attract quality teachers to relocate to hard-to-staff schools (Fowler, 2003; Prince, 2002). Our paper contributes to this literature by showing that the DT program can work like a self-administered training that increases the quality of local teachers and avoids the cost of relocating teachers. This finding is particularly important in light of the mixed results regarding the impacts of formal training program on teacher and student outcomes. For example, Loyalka et al. (2019) find that the government-funded National Teacher Training Program in China that cost at least $1 billion failed to improve teacher and student outcomes because it is overly theoretical in content and rote in delivery. By contrast, our findings suggest that the DT program is a very cost-effective, self-administered, and topic-specific teacher training that contributes more than one-third of the DT program induced improvements in student performance. To the best of our knowledge, the current study is the first to explore the role of a CAI program as a teacher training program in affecting student performance.

The remainder of the paper is organized as follows. Section 2 describes the background of the DT program and data used in our empirical analyses. Section 3 develops a theoretical model to guide the empirical analysis. Section 4 presents the empirical results about the effects of the DT program. Section 5 further explores variations in program exposure intensity. Section 6 offers some concluding remarks.

2. Background information about the DT program and our data

The DT program is a computer-assisted instruction program designed to close the vast cross-region gaps in educational resources and outcomes in China. It is a collaborative effort by three main parties: the Youchange China Social Entrepreneur Foundation (YCSEF), National Basic Education Resources Development and Sharing Alliance, and the Middle School Affiliated to Renmin University of China (Renmin MS). During the initial experimental stage of the DT program in 2013, it was run as a distance learning course on grades 7-9 math – students from 13 schools in remote areas attended, via the internet, live math lectures at Renmin MS, one of the best middle schools in China. After each live lecture, the local teacher reviewed the lecture's key points, answered any questions the students might have, assigned and graded homework, and offered one-on-one tutoring, if needed. The live-lecture model soon proved to be ineffective because in part the lectures were tailored toward the Renmin MS students who, in terms of preparation and ability, are far above their counterparts in the remote schools. In consultation with the local school principals, the sponsors of the DT program made adjustment in the second year – instead of live lectures the participating schools were granted access to recorded lecture videos online along with other materials, such as teaching notes, homework, exams, and solutions. Local schools and teachers had the discretion on how to integrate these materials into their instructions. By 2016 the DT program expanded to cover more than 200 schools in 21 provinces and covered other subjects besides math at the middle school level.

YCSEF initiated an evaluation study of the DT program with the goals of assessing the effect of the program on student learning outcomes and channels of influence. With these objectives in mind, they conducted a quasi-experiment focusing on a cohort of students from 9 representative participating middle schools.[6] Specifically, the 2014 cohort of incoming students from these middle schools were assigned to classes of 40 to 50 students each, as they normally would be, and a subset of the classes were then assigned as treatment and control classes or groups.[7] In each school there were at least two treatment classes and one control class. At least two treatment classes were taught by the same math teacher who had access to the DT program materials. While the teacher taught the same contents to the treatment classes, she was allowed to show lecture videos to only one of the classes, not the other. The class that got to watch the

[6] There were 10 schools selected for the evaluation study. One school was dropped from our study because it failed to provide information on student performance prior to entering middle school, which is the pre-treatment benchmark essential for our study.

[7] In this paper, we use class and group interchangeably to refer to students with the same treatment status.

lecture videos was therefore exposed to the DT program both directly -- receiving lectures by the high-quality teacher at Renmin MS -- and indirectly through the local teacher who prepared their lectures using the recorded lecture videos and other materials. We call this class full treatment class and the class that did not get to watch the lecture videos in class partial treatment class. The control classes were taught by a math teacher who had no access to any of the resources associated with the DT program and therefore were not exposed to the DT program directly or indirectly. The treatment status of a class as well as their local math teacher remained the same as the cohort of students progressed through the three-year middle school education. Under this research design, the difference in the pre- and post-treatment differences between treatment and control classes' performance would be attributable to the DT program. Similarly, the difference in the pre- and post-treatment differences between the full and partial treatment classes would be an estimate for the effect of direct exposure to the DT program. The estimated effect of indirect exposure can be similarly derived by comparing student performance in the partial treatment and control classes. These inferences would be causal if the assignment of treatment and control status is random. However, by the time we were invited to take part in the evaluation study in the winter of 2014, the experiment had been in progress for about one semester. Knowing the potential non-random selection issue, we made great efforts to gain insight on the process of treatment versus control assignment as part of the surveys we conducted in 2015 and 2016.

While in all participating schools the administrative board made the assignment decisions – classes and math teachers to take part in the evaluation study of the DT program -- prior to the start of the 2014-15 academic year, the method used varied across schools. A piece of critical information that we obtained through the survey of school principals is the initial assignment of incoming students to classes, which is independent of the DT program. Among the nine schools under study, seven reported to randomly assign incoming students into classes of 40 to 50 each. The other two schools, however, created honors classes and regular classes. The former consists of students with relatively higher academic achievement in terms the standard middle school entrance exam score. The regular classes were randomly formed with the remaining students. For schools with random initial assignment of classes, the assignment of DT treatment status is by extension also random. The only remaining concern is whether the assignment of the math teacher was random. Again, according to the survey of school principals, four schools made the teacher selection randomly, two selected their average-quality math teachers, and three selected their best math teachers to teach the treatment classes. However, a closer inspection of the distribution of math teachers in terms of professional title, an indicator of quality and experience of the teacher, shows that there is no tendency of assigning higher quality teachers to treatment classes. Among the 12 math teachers who taught the treatment classes, 3 held the advanced title, 7 the intermediate title, and 2 the elementary title. The corresponding title distribution of the 13 math teachers assigned the control classes is 6, 6, and 1. It seems that we can rule out at the outset the possibility that any positive impact of the DT program was due to a disproportionate of higher quality teachers being assigned to the treatment classes. It is important to note that we control for teacher's quality in all our regression analyses. We also conduct robust checks to assess the extent to which our main results are sensitive to the non-random assignment of teachers in a subset of our sample schools.

We conducted surveys of students and teachers. The student survey focused primarily on student's demographic and family characteristics. The teacher survey collected information on teacher's professional qualification and usage of the teaching materials provided by the DT program, especially the usage of the lecture videos during lesson preparation and in class. We also obtained from each school the student's test scores on three subjects: math, Chinese, and English. There are three sets of test scores. The first are test scores of the middle school entrance examination administered at the end of the elementary school education by each school district in which elementary schools are feeder schools of their middle schools. These are the only comparable performance indicators prior to the middle school education and will be used as the pre-treatment benchmarks in our empirical investigation. The second set of performance measure consists of the end-of-semester final exam scores.

The third performance measure is the high school entrance exam score. We choose the high school entrance exam math score rather than the semester-end final exam score as our main post-treatment performance measure for two reasons. First, the high school entrance exam is a standardized test administered by the provincial bureau of education to all middle school graduates in the province. Second, the tests are written and graded by educational experts based on the middle school curriculum and without the involvement of individual schools or teachers. In contrast, the semester-end exams are written and graded by an individual teacher or a group of teachers in the school and therefore the level of difficulty and grading of the exams may even vary across classes within a school.

Table 1 presents the basic information about the distribution of students among the three types of classes, and summary statistics on the usage of lecture videos as well as student and teacher characteristics. Of the 1887 students from 43 classes included in the evaluation study, 30% are in the full treatment classes, 29% in the partial treatment classes, and the remaining 41% in the control classes. The average class size is 44 students. There are 25 math teachers participated in the evaluation study. Twelve of these teachers taught the full and partial treatment classes. They spent an average of 6.42 hours per week watching the lecture videos while preparing for their own lectures and devoted an average of 1.27 hours of class time per week to showing selected parts of the lecture videos. Given a typical weekly class schedule in middle school includes five 45-minute sessions of math, these statistics indicate that the local math teachers on average went over the lecture videos more than once and the showing of the selected lecture videos occupied one-third of instruction time of the full treatment classes. The latter indicates that the use of the lecture videos is substantial as a share of the weekly instruction time. The gender distribution of students is even at 50% of male and female each. A slightly larger proportion of the students were from rural areas than from urban areas, 58% versus 42%. This is expected because the DT program targeted schools in remote locations where a large share of the population is rural. Our main measure of student family background is educational attainment of parents. The average years of schooling of father is 8.4, about one year higher than mother's 7.1. There are three levels of professional ranking of teachers: advanced, intermediate and elementary. These titles are usually awarded according to professional qualification and training, teaching experience and performance. About one-third of the math teachers have the advanced title, 55% of them hold the intermediate title, and the remaining 14% hold the elementary title.

Table 2 presents the average test scores in math, Chinese and English by treatment status for both the pre-treatment and post-treatment periods. As noted, the scores are standardized by school and period and therefore reflect relative performance. The average pre-treatment scores are all close to zero across three groups of students and three subjects, indicating comparable initial performance of students of different treatment status. Same pattern is observed for the average post-treatment scores in Chinese and English, but not in math, which is the target of the

DT program. The average post-treatment math scores diverged across treatment status. While it is more than half standard deviation above the mean and near the mean for students in the full and partial treatment classes, respectively, the average math score for students in the control class is about 0.4 standard deviation below the mean. The lower part of table 2 shows the estimated cross-group difference in test scores along with standard errors. The differences in the pre-treatment test scores in all three subjects (including math) are small and not statistically significant. In contrast, the differences in the post-treatment math score are not only statistically significant at the 1% level but substantial in magnitude, consistent with the presence of positive impacts of the DT program. It should be noted that the cross-group differences in the post-treatment scores in Chinese and English remain statistically insignificant.

3. Theoretical Framework

The DT program leveraged on web-based technologies can improve student performance through two channels. The first is that it enables students in the remote school to be taught by a high-quality teacher from an elite urban school -- the local teacher shows selected parts of the lecture videos in class to complement their own lectures and other instructional activities. The second is that it helps improve the instruction quality of the local teacher, who watches the lecture videos during lesson preparation. To the local teacher, the DT program works like a web-based training that they use to gain new skills or human capital.

Following Acemoglu et al. (2014), we assume that the human capital of all students before they enter middle school is the same within a school and the human capital of all teachers within a school is the same. They are denoted by e_r and h_r, respectively, for the remote school. For the elite urban school, they are denoted by e_u and h_u. Student human capital endowment comes from prior education and family endowment. We also assume that both student human capital and teacher human capital are higher in the elite urban school than in the remote school. That is $e_u>e_r$ and $h_u>h_r$. In the absence of the DT program, the post-schooling human capital of the student in the remote school, y_r, is expressed as a function of student human capital endowment and teacher human capital:

$$y_r = e_r^{1-\alpha} h_r^{\alpha}, \qquad (3.1)$$

where $\alpha \in (0,1)$. Note, h_r can be thought of as an aggregator of skills of the teacher allocated across different teaching tasks, such as lecturing, class discussions, one-on-one interactions and so on. Rewrite equation (3.1) in logs, we have

$$lny_r = (1-\alpha)lne_r + \alpha lnh_r. \qquad (3.2)$$

With the DT program, students and teachers in the remote school gain access to the lecture videos of the urban teacher teaching their classes. Therefore, the urban teacher enters the human capital production function both directly through lecturing and indirectly through improving local teacher's human capital. The post-schooling (or post-treatment) human capital $y_r^{'}$ is given by:

$$y_r^{'} = e_r^{1-\alpha}[(\varphi h_u)^{\beta}(h_r^{'})^{1-\beta}]^{\alpha}, \qquad (3.3)$$

where $0 < \varphi < 1$ is an efficiency indicator, $0 \leq \beta \leq 1$ is the share of teaching services (lecturing only) provided by the urban teacher and $1-\beta$ is the share of lecturing and other services provided by the local teacher, and all other variables are as defined before. This formulation is inspired by Acemoglu et al. (2014). However, there are two critical modifications based on what we learned about the DT program. The first is that we assume that lecturing is done by both the urban high-quality teacher and the local teacher. The latter is also responsible for other teaching services. This is in contrast to Acemoglu et al. (2014) who assume that a "superstar" teacher does all the lecturing via the web-based technologies and a local teacher provides all other teaching services. The main consideration for our shared lecturing assumption is that the urban teacher's lectures are prepared based on the ability or human capital of urban students (e_u), which is greater than the human capital endowment of students in the remote school. It is conceivable that low-ability students would have difficulty in following the lectures that are tailored to high-ability students. In fact, as noted in the previous section, the DT program was originally meant to bring lectures live to classes in the remote school via the Internet. That model proved ineffective because of large gaps in academic background between students in the remote school and students attending one of the best middle schools in China. This is also a main reason why the local teacher chooses to show only parts of the lecture videos in class. We introduce the efficiency indicator φ to capture the fact that the urban teacher substitutes the local teacher for only a fraction of the lecturing service. We further assume that the share of the urban teacher's lecturing that is an effective input for low-ability students depends on the ability or human capital gap between students in the urban and remote schools, i.e., $\varphi = e_r/e_u$ --the larger the gap the smaller the share of the lecture videos that is suitable for students in the remote school.

The second modification we introduce relates to the role of the local teacher. In Acemoglu et al. (2014), a "superstar" teacher does all the lecturing remotely and frees the local teacher up from lecturing so that they can devote all their time to other complementary teaching services. As a result, more student human capital is produced not just because of better quality lecturing but also because of increased non-lecturing services provided by the local teacher. In our model, we assume that the local teacher improves their teaching skills or accumulates new human capital through learning from the lecture videos during lesson preparation and there is no change in the allocation of the local teacher's time across different teaching activities, because the teacher is in the class monitoring and supervising students when lecture videos are shown. The post-DT program human capital of the local teacher can be expressed as a function of own human capital endowment and that of the urban teacher's, $h_r^{'} = f(h_r, h_u)$.

Rewrite equation (3.3) in logs, we have

$$lny_r^{'} = (1-\alpha)lne_r + \alpha\beta ln(\varphi h_u) + \alpha(1-\beta)lnh_r^{'}. (3.4)$$

From equations (3.2) and (3.4), we can compute the percentage increase in human capital of students in the remote school due to the DT program as

$$ln\frac{y_r^{'}}{y_r} = \alpha\beta ln\frac{\varphi h_u}{h_r} + \alpha(1-\beta)ln\frac{h_r^{'}}{h_r}, \qquad (3.5)$$

where the first term is the gain attributable to superior lecturing service by the urban teacher and the second term represents the gain attributable to increases in human capital of the local teacher. In the context of this evaluation study, equation (3.5) indicates the increase in human capital of students in the full treatment class relative to peers in the control class. When $\beta = 0$, equation (3.5) shows the increase in human capital of students in the partial treatment class relative to those in the control class.

4. Empirical model and results

We estimate the causal effect of the DT program using the following difference-in-differences (DD) regression model,

$$y_{ikjt} = \beta_0 + \beta_1 F_{ikj} + \beta_2 P_{ikj} + \delta T_t + \theta F_{ikj} T_t + \rho P_{ikj} T_t$$

$$+ \gamma \boldsymbol{X}_{ikj} + \alpha \boldsymbol{W}_{kj} + \boldsymbol{\omega}_j + u_{ikjt}, \quad (4.1)$$

where y_{ikjt} denotes the math test score of student i in class k, school j and period t, F_{ikj} is a dummy variable equaling 1 for students in the full treatment class and 0 for all others, P_{ikj} is a dummy variable equaling 1 for students in the partial treatment class and 0 for all others, T_t is a period dummy equaling 1 for periods that the DT program is in effect and 0 otherwise, $\boldsymbol{X}_{ikj}$ represents student characteristics, $\boldsymbol{W}_{kj}$ represents math teacher characteristics, $\boldsymbol{\omega}_j$ is school fixed effects, and u_{ikjt} is the error term. We estimate equation (4.1) using Generalized Least Squares to account for within-school-class correlation of students' test scores without imposing a specific distribution on the residuals. To ensure comparability of test scores across different semesters and schools, we follow the literature and convert the raw scores into standardized z-scores by school and semester so that the standardized score reflects each student's relative performance ranking in their grade and school (see, for example, Matsudaira, 2008). The pre-treatment (*T=0*) test scores come from the middle school entrance examination, which is a standardized assessment test at the end of primary school education and covers math, Chinese and English. As noted in section 2, this is the only pre-treatment performance measure that is available to us and is used as such in all our regression analyses. While we have the end-of-semester final exam scores for all six semesters throughout the three-year middle school education, we use the last middle school final scores as the main post-treatment performance measure (*T*=1). The last final is a standardized assessment test, also known as the high school entrance exam, which is administered by provincial bureau of education to all middle school graduates and covers math, Chinese, and English.[8] We use the rest of the end-of-semester final exams scores to provide supplemental evidence on progression and persistence of the impact of the DT program. We do not emphasize these interim performance measures because the treatment was ongoing.

4.1 Basic results

[8] Besides serving as measures of students' knowledge of the official curriculum, these scores also determine the types of high school a student can enroll in. High scores are required for attending high quality, well-funded high schools – the so-called region-level key high schools or province-level key schools.

Table 3 presents the estimates based on two versions of equation (4.1). We begin with the fixed-effect regression without any control variables in column 1. The DD estimates for the effects of the full and partial treatments are positive and statistically significant at the 1 percent level, suggesting that the DT program help improve math performance of students in the treatment classes. Specifically, the DT program increases math score of students in the full and partial treatment classes by 0.978 and 0.343 standard deviations on average, respectively. These effects are equivalent to moving a student initially ranked at the school median in a normal distribution up to the 84th percentile and 63th percentile, respectively. As expected, a larger improvement in math score is found for students in the full treatment class. This indicates that the use of lecture videos as part of the in-class instruction produces additional benefits beyond the gain in student performance due to improved instruction of the local math teacher, who uses the lecture videos to aid their lesson preparation. While it is tempting to infer from these DD estimates that the direct effect of the DT program (defined as the use of the selected lecture videos as part of the in-class instruction or student exposure to the DT program) is 0.635 (i.e., 0.978-0.343) and is greater than the indirect effect (defined as local teacher's use of the lecture videos for lesson preparation or teacher exposure to the DT program) of 0.346, the direct effect is likely to be smaller than the difference in the DD estimates. This is so because teacher exposure that improves local teacher's instruction quality may amplify the effect of student exposure. Since we do not have a group of students treated exclusively with only the lecture videos, we cannot empirically isolate the direct effect. In view of the potential presence of such complementary relationship, it is sensible to interpret the difference between the DD estimates as an upper-bound estimate for the direct effect of the DT program.[9]

The estimates associated with the full and partial treatment indicators are statistically indistinguishable from zero, suggesting that the pre-treatment math scores are the same on average among students in the treatment and control classes. The estimate on the treatment period indicator T is negative and statistically significant, suggesting that students in the control group see their math scores decrease by 0.396 standard deviations on average over the course of the three-year middle school education. It should be emphasized that, since the math score is

[9] In section 5 we implement a variant of the regression model that allows for teacher exposure to affect the impact of student exposure. We find no statistical evidence consistent with the presence of complementary relationship.

standardized by school, this negative estimate indicates a fall in the relative ranking, not necessarily in absolute math score, of students in the control class.

In model 2 we include several covariates that may correlate with student academic performance. The covariates are student gender, hukou status, parental education in years of schooling, and math teacher's professional title.[10] As the estimates in column 2 indicate, the inclusion of these control variables has virtually no effect on the DD estimates and the estimates associated with the treatment and treatment period dummies. These indicate that the treatment status and impact of the treatment are uncorrelated with student and teacher characteristics. In other words, the treatment effects are not likely coming from differences in student and teacher backgrounds between the treatment and control groups. Three of the added covariates obtain statistically significant coefficient estimates. Both father's and mother's schooling in years are positively associated with math test score. Students taught by teachers with the intermediate title receive higher scores than students taught by teachers with the elementary title (the benchmark category). It is interesting to note that students taught by teachers with the advanced title do not statistically outperform students taught by teachers with the elementary title.

As a specification test and a check on if the estimated treatment effects are spurious, we repeat the fixed-effects regressions of columns 1 and 2 using the corresponding standardized test scores in Chinese and English. Since the DT program is designed to improve student performance in math, it is unlikely to help increase student performance in other subjects. While we cannot completely rule out the possibility of spillover effects that students treated with the DT program for math may become more motivated to do well in other subjects they are learning in school, the presence of treatment effects on Chinese and English test scores would call into question whether the DD estimates based on math test scores can be interpreted as causal. As the estimates reported in columns 3 and 4 for Chinese and columns 5 and 6 for English show, the DT program has absolutely no statistically meaningful impacts on Chinese and English test scores for students in both the full and partial treatment classes. There might be a hint of spillover effects as some of the DD estimates are positive. However, given the large standard errors, we cannot reject the null hypothesis that the DT program has no impact on student performance in

[10] Liu (2005) contains a discussion on the Hukou system and its role in determining the rural-urban disparities in China.

Chinese and English. These findings are reassuring that the positive treatment effects of the DT program on math performance are unlikely spurious.

4.2 Randomness of the treatment assignment

One assumption underlying the causal inference in the DD framework is random assignment of treatment status. This is especially critical since we do not have information on student performance over multiple periods prior to the treatment, which would allow us to check the presence of parallel trends between the treatment and control classes. However, pre-treatment trends would not be a concern if the assignment is random. As noted in the preceding section, the treatment assignment for the purpose of evaluating the effectiveness of the DT program may not appear completely random. In this subsection we present some statistical evidence to show that the assignment is largely consistent with what we would expect if the assignment was randomly determined.

We first conduct balance tests across the full treatment, partial treatment and control groups in terms of the pre-treatment math score. As noted early, the test scores come from a standardized assessment test administered to all primary schoolers at the end of the primary school education. If the treatment assignment is not systematically correlated with the pre-treatment math score, the assignment would be as if done through a random process. Also, since the local math teacher plays an important role in implementing the DT program and teacher ability and experience are important determinates of student performance, the assignment of local teachers to treatment and control groups should ideally be random as well. Teacher's assignments would be approximately random if the probability of having a math teacher with a particular professional rank (a proxy for ability and experience) does not vary for students across the treatment and control groups.

Column 1 of table 4 presents the balance test for the pre-treatment math score. The estimates are obtained from a school fixed-effects regression without any other covariates and indicate the within-school differences in the test score between the full treatment and control groups, and between partial treatment and control groups. As the estimates for the treatment indicators are not statistically significant, there is no difference in the pre-treatment math score across the three groups of students. Therefore, the pre-treatment math score does not appear to be a criterion for treatment assignment. Furthermore, if treatment assignment is analogous to random we would expect a balanced distribution of students in terms of performance in other subjects as well. In columns 2 and 3, we repeat the regression of column 1 with the pre-treatment test scores in Chinese and English, respectively. The estimates on the treatment indicators are not statistically significant, suggesting no systematic difference in the pre-treatment performance in Chinese and English between the treatment and control groups. It seems safe to conclude that the pre-treatment performance of students in math or any other subject plays no role in treatment assignment. As for the assignment of the math teacher, none of the estimates on the treatment indicators reported in columns 4 and 5 is statistically significant, suggesting that students in the full and partial treatment groups have statistically equal chances as students in the control group to have a math teacher with the advanced or intermediate title vs elementary title. We also check whether the pre-treatment math score is correlated with teacher characteristics as an indirect test for random assignment of teachers to treatment groups. If the assignment is approximately random, teacher ability would be uncorrelated with the pre-treatment math score, which, as a performance indicator of students at the end of primary school education, should have nothing to do with the quality of the middle school math teacher if assignments of students and teachers are random. In column 6 we expand the regression model of column 1 to include teacher's professional title and student characteristics as covariates. The estimates for the effects of teacher's title are not statistically significant. Taken together, these balance tests suggest that the assignments of students and math teachers are consistent with the pattern of a random assignment.

In columns 7 through 10 of table 4 we test whether student characteristics are balanced across the treatment and control groups, as would be expected in an experiment with random assignment. While we do not find any statistically significant difference in the distribution of student gender and hukou status, students in the treatment groups tend to have parents with more years of schooling comparing to their counterparts in the control group. Would this cause any bias in the estimated effects of the DT program? The results reported in Table 3 suggest that the answer is no. As noted previously, the DD estimates for the effects of the DT program are virtually insensitive to the inclusion or exclusion of parental education as covariates, suggesting that treatment assignment is independent of parental education background. The estimated coefficients on father's and mother's education both are positive and statistically significant at the one percent level indicating positive impacts of parental education on children's performance, as documented by numerous studies. But the insensitivity of our DD estimates to the inclusion of parental education variables suggests that parental education does not augment or diminish the impacts of the DT program on the treated students. In the next sub-section, we will further address this imbalance issue by assessing the potential selection bias it may cause.

4.3 Robustness checks

As a way of checking the sensitivity of our main results to imbalanced distribution of parental education across the treatment and control groups, we repeat our baseline regression (model 2 of table 3) using trimmed samples with more balanced parental education. A closer inspection of parental education in each of the three groups reveals that the distribution of parental education (for both parents) is skewed to the right for the treatment groups but to the left for the control group. We trim our sample by parental education to construct three subsamples. For the first subsample we remove from the baseline (full) sample students in the treatment groups whose father or mother received post-secondary education. For the second subsample we remove from the baseline sample students in the control group whose father or mother received no formal education. The third subsample is the intersection of the first and second subsamples. While the average parental education remains somewhat higher for the treatment groups than for the control group in the first subsample, the reverse is true in the second and third subsamples. On the whole, comparing to the baseline sample, the cross-group distribution of parental education is obviously more balanced in each of the trimmed subsamples. Columns 1 through 3 of table 5 present the DD estimates based on the three subsamples, respectively. First, all the estimates (both for the full and partial treatment effects) are positive and statistically significant, suggesting that the treatment effects we have identified are not driven by imbalanced distribution of parental education across groups. Second, none of these six estimates is statistically different from their counterparts based on the baseline sample. This suggests that potential selection bias arising from students with better family background being selected into the treatment groups is statistically negligible. Third, since all of the six estimates are quantitatively larger than their counterparts based on the baseline sample, selection bias associated with parental education, if exists, would be in the negative direction, i.e., against finding treatment effects.[11]

[11] We also repeat regressions for Chinese and English test scores using these subsamples and obtain DD estimates that are not statistically significant, consistent with the results reported in columns 4 and 6 of Table 3.

As an alternative way of testing the sensitivity of our main results to imbalanced parental education background, we expand the baseline model of column 2 of table 3 to include interaction terms of F*T and P*T with father's and mother's education. These interaction terms allow the treatment effects to vary with parental education. The DD estimates from this regression specification are reported in column 4 of table 5. Similar to the previous three cases we just discussed, the DD estimates remain positive and statistically significant but are larger in magnitude than the baseline estimates. All the newly added interaction terms obtain negative coefficient estimates (not reported in the table), suggesting that the treatment effects diminish with parental education and that the baseline estimates may be biased toward zero because they are based on the sample in which students in the treatment groups are more likely to come from families with better educated parents. However, such downward bias would be negligible because none of the interaction terms involving parental education is statistically significant.

Although the balance tests indicate that treatment assignment in our sample mimics a random experiment, we conduct additional robustness checks on the sensitivity of our baseline results to potential selection bias arising from non-random assignment of students and teachers to treatment versus control groups. Recall, seven of the nine schools in our sample randomly assigned the incoming students into classes prior to the DT program assignment. Since the initial class formation is random, the selection of classes for full and partial treatment can be considered random as well. The only concern would be if the selection of the math teacher is also random or not. Fortunately, this issue can be largely addressed by including teacher characteristics as controls in the regression model, as we have done so far. Column 5 of table 5 presents the DD estimates based on the baseline regression of column 2 of table 3 and the subsample of schools that made the initial class assignment in a random fashion. Column 6 presents the DD estimates from the baseline regression using the subsample of schools that randomly selected the math teacher for the treatment classes, while column 7 contains the DD estimates of the same regression using the subsample of schools that randomly assigned both class and math teacher the treatment status. As shown in the last three columns of table 5, the DD estimates for the effect of the DT program on the full treatment group range from 0.960 to 1.113 and are statistically significant at the 1% or 5% level. Since treatment assignment is random or controlled for (in column 5), these estimates are free of selection bias. Furthermore, the fact that these estimates are remarkably comparable to our baseline DD estimate of 0.978 suggests that selection bias associated with non-random assignment, if exists, is not substantive enough to invalidate our inference concerning the effects of the DT program. In fact, we cannot reject the null hypothesis that the DD estimates for the effect of full treatment reported in columns 5 through 7 are equal to our baseline DD estimate. The same can be said about the DD estimates for the effect of partial treatment – they are very comparable to the baseline estimate of 0.343, in terms of both magnitude and statistical significance. We also implement these regressions with Chinese and English test scores as the outcome variables and find no positive and statistically significant treatment effects, consistent with our findings based on the full sample.

It is worth noting that the estimated effect of direct student exposure to the DT program – the difference between the DD estimates for the full and partial treatment classes—is less susceptible to non-random assignment of instructors because the treatment classes are taught by the same math teacher. The difference in these two DD estimates is always positive and statistically different from zero based on the full sample (table 3) and various subsamples (table 5). This result also holds in the related regression analyses that follow.

4.4 Accounting for pre-treatment trends

The key identifying assumption underlying the DD regression when treatment assignment is not random is that in the absence of the DT program the performance difference between students in the treatment and control groups would be constant over time. This common trend assumption can be investigated using data from multiple pre-treatment periods. However, this is not possible for us because we have information on student performance for just one period prior to the treatment. We could dismiss any concern about the parallel trend assumption by pointing to the statistical evidence we presented early that indicates that treatment assignment in our sample mimics an experiment with random assignment. To add credence to our main findings we conduct additional empirical tests and robustness checks to account for possible non-parallel trends in math test score between the treatment and control groups that would have persisted throughout the treatment period without the DT program. One way to account for the presence of different trends is to add, as a control variable, a performance indicator that is highly correlated with math test score in the absent of the DT program but is itself unaffected by the DT program. We think that test scores in Chinese and English satisfy both requirements. On average, students who excel in one subject are more likely than not to do well in other subjects as well. Using the pre-treatment test scores in math, Chinese and English, we find that they are highly correlated with each other. The pair-wise correlation coefficients are: 0.513 (between math and Chinese), 0.631 (between math and English), and 0.692 (between Chinese and English).[12] As we showed early (columns 3 through 6 of table 3), the DT program, which again is designed to improve student performance in math, has no effect on Chinese and English test scores of the treated students relative to those of students in the control group.

We run two variants of our baseline DD regression model. In the first, we include Chinese test score or English test score or the average of the two as well as their interaction terms with the treatment indicators as additional regressors to account for group-specific trend in math performance that would have transpired in the absence of the DT program. The DD estimates from these specifications are reported in columns 1 through 3 of table 6. The estimates for the effect on math score of students in the full treatment group range from 0.934 to 0.977 and are not statistically different from our baseline estimate of 0.978. Similarly, the estimates for the effect on math score of students in the partial treatment group range from 0.337 and 0.380 and are comparable to the baseline estimate of 0.343.

The second variant of our DD regression specification is a form of triple-difference model, where we account for group-specific counterfactual trend in math performance by taking the difference between math and Chinese test scores, or between math and English test scores, or between math test scores and the average of Chinese and English test scores and use the difference as the dependent variable. Essentially, we assume that in the absence of the DT program the cross-group difference in math score, if exists, would mirror the cross-group difference in Chinese (or English or the average of the two) score that is unaffected by the DT program. In the context of a triple-difference regression model, the Chinese test score (or English or both Chinese and English score) plays the role of accounting for time-varying confounders that develop differently across the treatment and control groups. In this regression specification, positive treatment effects on performance in math would show up as larger increases in the test

[12] It is interesting to note that the corresponding correlation coefficients based on the post-treatment test scores are: 0.397, 0.400, and 0.758. The weakening of the correlations between math and Chinese and between math and English is consistent with our findings that the DT program improved math performance of students in the treatment group but had no effects on their performance in Chinese or English. The correlation between Chinese and English, two subjects unaffected by the DT program, remains high, consistent with the notion that students excel in one subject are likely to stand out in other subjects as well.

score differentials (between math and Chinese, for example) for students in the treatment groups than for students in the control group.[13] This is exactly what we find. The DDD estimates reported in columns 4 through 6 are all statistically significant at the one percent level and are comparable to the baseline estimates. In fact, they are statistically the same as the baseline estimates.

Continuing the line of reasoning that student performance in all three subjects would progress in tandem in the absence of the DT program. Under this assumption we can detect a treatment effect of the DT program from a DD regression model using test scores in a non-math subject (the control subject) as the counterfactual trend for math (the treatment subject) scores. Table 7 presents the DD estimates from such regressions separately for our full treatment, partial treatment, and control groups, with Chinese (column 1), English (column 2), and the average of the two (column 3) as the control subject, respectively. The estimates in panel A suggest that the DT program improve math test scores of students in the full treatment group by 0.515 standard deviations on average relative to their Chinese test scores. The improvements relative to English and the average of Chinese and English test scores are 0.566 and 0.563, respectively. As the estimates in panel B indicate, math performance of students in the partial treatment group did not deviate from their performance in Chinese or English or the average of the two subjects. But students in this group did better than students in the control group, whose performance in math, as the estimates in panel C suggest, fell relative to their performance in Chinese, English, and the average of the two subjects by 0.361, 0.401, and 0.388 standard deviations, respectively. Students in the partial treatment group benefited from the DT program because if they were not treated at all their relative performance in math would have been lower, similar to what happened to students in the control group. It should be noted that the DD estimates in table 7 are not directly comparable to our baseline estimates which assume the control group's performance in math as the counterfactual for the performance of the treated students in the absence of the DT program. However, since the estimated coefficients on T (the treatment period indicator) suggest that there is no significant upward or downward trend in student performance in Chinese or English for all three groups, differences in the DD estimates between the full treatment (panel A) and control (panel C) groups and between the partial treatment (panel B) and control (panel C) groups approximate our baseline DD estimates. For example, the estimates in column 1 produce

[13] These are the triple-difference estimates.

the effect of full treatment of 0.876 (0.515-(-0.361)) standard deviations and the effect of partial treatment of 0.330 (-0.0307-(-0.361)) standard deviations. These derived estimates of the effects of the DT program are in the ballpark as our baseline estimates of 0.978 and 0.343, respectively.

4.5 Dynamics of the treatment effects

Thus far the post-treatment performance measure we use is the end-of-middle-school assessment test score in math. One may be concerned that the treatment period of six semesters or three years is too long to rule out the existence of confounding factors that might have been responsible for the performance improvement of the treated students. To address this issue, we use the semester-end math exam score as the post-treatment outcome variable and repeat the baseline regression for each of the six semesters. This empirical strategy allows us to estimate the treatment effects after a treatment of various lengths, ranging from one to six semesters. As before, the pre-treatment performance measure is the math score from the end-of-primary-school assessment exam. The semester-by-semester DD estimates are reported in table 8. There are three noteworthy points about these estimates. First, the positive impacts of the DT program are evident at the end of the very first semester -- students in the full treatment group saw their math scores improved by 0.427 standard deviations on average, while students in the partial treatment group experienced an improvement of 0.104 standard deviations. Although the effect of the partial treatment is not statistically significant at the end of the first semester, it becomes significant at the 10 percent level by the end of the second semester. The positive impacts of the DT program are appreciable within a short period of treatment. Second, the positive impacts of the DT program are cumulative over the course of the three-year middle school education. This is particularly true of the effect of the full treatment--the estimate for the effect of the full treatment increases over time, except for a small dip at the end of the fourth semester. In other words, five out of six semester-on-semester changes in math score are positive. This pattern of incremental and sustained improvement adds credibility to our main findings. Third, the impacts of the DT program strengthen during the last two semesters in terms of both magnitude and statistical significance. About 45% of the overall performance improvement (0.978 standard deviations) of students in the full treatment group occurred in the last two semesters. The corresponding number is 49% for students in the partial treatment group.

The persistence of the impacts of the DT program is also corroborated by regression analysis based on the pooled sample that includes test scores from all six semesters plus the pre-treatment test score. In this regression, we replace treatment period indicator, T, with a time trend, Time (a whole number from 0 to 6), and F*T and P*T with F*Time and P*Time, respectively. The coefficients on the latter two interaction terms provide estimated effects of full and partial treatments on the time trend of the math score. Positive estimated coefficients would indicate that the DT program has resulted in a steeper trajectory of math scores for the treatment classes relative to the control class. As the estimates reported in column 1 of table 9 indicate, this is exactly what we find. The DT program increases the math score of students in the full treatment class by an average of 0.138 standard deviations per semester and the math score of students in the partial treatment class by an average of 0.0513 standard deviations per semester. The negative and significant estimate for the time trend suggests that students in the control class saw their performance in math fell by an average of 0.0568 standard deviations per semester. Again, since the scores are standardized by school and semester these estimates imply the impacts of the DT program on the relative, not absolute, performance in math across the three groups of students. We also repeat this panel regression with Chinese and English test scores and find no such effects on trend that are attributable to the DT program, as we would expect since again the DT program aims at improving student performance in math (see the estimates reported in columns 2 and 3 of table 9).

4.6 Placebo tests

To rule out the possibility of our estimated treatment effects being an artifact of multiple hypothesis testing, we also conduct a nonparametric permutation test using the *p*-values from our baseline regression (i.e., model 2 of Table 3) as critical values. We generate 2000 placebo samples in which we randomly reassign full-treatment, partial-treatment, and control status to students within each school. Using each placebo sample, we replicate model 2 of Table 3. We then calculate the fraction of placebo simulations in which the *p*-values for the DD estimates are smaller than the corresponding true *p*-values for the estimates in the actual treatment assignments. We find none, which means that fewer than 1 percent of the placebo replications produce *p*-values that fall below the values we estimate. Therefore, the permutation test generates a *p*-value of less than 1 percent for the null hypothesis that the DT program has no effect on student performance in math.

We also perform similar permutation tests for the treatment effects on Chinese and English test scores, respectively. Since the treatment effect estimates based on the actual assignments are not statistically different from zero, we compute the fraction of placebo simulations in which the *p*-values for the DD estimates are less than 0.01. We find that out of the 2000 placebo simulations only 2 and 7 samples produce the *p*-values smaller than 0.01, respectively for Chinese and English scores. This means that the probabilities that we falsely reject the null hypothesis that the DT program has no effect on Chinese and English scores are less than 1 percent.

4.7 Effect heterogeneity

As we noted early, the effects of the DT program do not seem to vary by student and teacher characteristics. To provide further and direct evidence we next estimate our baseline model using subsamples created based on student or teacher characteristics. Panels A and B of table 10 present the DD estimates by student gender and by student *hukou* status, respectively. While male students benefit slightly more than female students in the full treatment group and the reverse is true in the partial treatment group, the differences are quantitatively small and statistically insignificant. The DD estimates suggest that students with rural hukou benefit more from the DT program than students with urban hukou. However, the difference is statistically significant at the 10 percent level only for the effect of full treatment.

Panels C and D present the DD estimates by parental educational attainment. We divide our sample into two subsamples by if a parent finished middle school. The effects of full treatment are very similar across the two subsamples and the difference is not statistically significant. But the effect of partial treatment is larger for students with parents who completed middle school education or more, and the effect is significantly larger for students whose father completed middle school education or more. Panel E presents the DD estimates by teacher's professional rank. Although the estimated effect of full treatment increases with teacher's professional rank, none of the cross-rank differences is statistically significant. We do find significant heterogeneity in the effect of partial treatment--students taught by a math teacher with intermediate title benefit from the DT program significantly more than other students in the partial treatment group.

To investigate if the effects of the DT program vary across students of different abilities, we divide our sample into two subsamples at the median of the pre-treatment math scores. It is worth noting that the shares of students from treatment and control groups in the two subsamples are remarkably similar to that of the whole sample. Among students in the lower-50% subsample, 30.3% come from the full treatment group, 28.4% the partial treatment group, and 41.3% the control group. The corresponding numbers for students in the upper-50% subsample are 30.5%, 29.2%, and 40.3%. These distributions are very comparable to the whole sample's 30.4%, 28.8%, and 40.9%. The DD estimates reported in panel F suggest that low performing students benefit from exposure to full treatment more than high performing students, but the difference in the estimates is not statistically significant. The estimates for the effects of partial treatment are not statistically different across the two subsamples, suggesting that high and low performing students in the partial treatment group benefit equally from the DT program.

4.7 Student dropout, transfers, and spillovers

One may be concerned that our main results could be biased upwardly or even largely driven by student dropout or between-class transfers if low performing students in the treatment classes dropped out of school or transferred to the control class and better performing students in the control class transferred to the treatment classes. There are several pieces of evidence that help dispel this concern. First, middle school education is mandatory under China's nine-year compulsory education law. While dropouts still occur in rural middle school, only a small fraction of students fail to complete middle school education and it is hard to imagine that poor performance in math is the primary reason for dropping out of school. Second, in each of the schools included in our sample the size of the classes of different treatment status is nearly the same with virtually no change throughout the three-year period under study, suggesting rare to no incident of student dropout. Third, to the extent that student performances in different subjects are highly correlated as we showed early, the highly comparable Chinese and English scores across the treatment and control classes both prior to and post the treatment suggest that dropouts or transfers, if there were any at all, would not have occurred in a way that results in a greater proportion of high-performing students ending up in the treatment classes and a greater proportion of low-performing students ending up in the control class. Fourth, and more important, neither school principals nor local teachers noted student dropout as a serious problem during the surveys and the schools were instructed from the outset of the evaluation study not to change treatment status of students and teachers throughout the three-year middle school education.

Another concern is the presence of spillovers that may occur through interactions between students from the treatment and control classes and between teachers of the treatment and control classes. However, to the extent that such spillovers are expected to improve math scores of the students in the control class, the presence of spillovers would lead to underestimation of the effects of the DT program. Therefore, the impacts of the DT program on student performance are potentially larger than what we reported in this study.

5. Exploring variation in treatment intensity

A key component of the DT program is to grant the local math teacher the access to the recorded lecture videos from an elite urban middle school. As far as we know, the program does not set any guideline on how the teacher should use this particular resource. As such, it is entirely up to the local math teacher how many hours to spend on watching the lecture videos during lesson preparation and how many hours and which parts of the lecture videos to show directly to students in class. Since a typical middle school class schedule includes one 45-minute session of math lecture daily, there are 3.75 hours of new recorded lectures each week made available to the local math teacher. As shown in table 1, the local math teacher who teaches students in the treatment classes spent between 5.25 and 7.25 hours per week watching the lecture videos when preparing for their lessons, which suggest that the local math teacher has not only gone through the entire recorded lectures but watched some parts more than once. Through this process the local teacher also selects parts of the lecture videos to show to students in class. The selected segments may be something that is hard for the teacher to incorporate into their instructions, such as instructions/experiments involving certain instruments that are not available in the local school. The selected parts may also be something more appropriate for the students in terms of level of difficulty. Our sample students who are in the full treatment group spent between 1 and 1.5 hours per week watching parts of the lecture videos in class, which amounts to one to two sessions or up to 40% of the weekly class periods for math. Since there are variations in the amount of time the teacher as well as students spent on watching the lecture videos, students in the treatment groups are exposed to the DT program with different intensity both directly through watching parts of the lecture videos in class and indirectly through improved instruction quality of the local teacher who benefits from watching the lecture videos. To investigate if treatment intensity matters, we estimate the following modified DD regression model:

$$y_{ikjt} = \beta_0 + \beta_1 F_{ikj} + \beta_2 P_{ikj} + \delta T_t + \theta SW_{ikj}T_t + \rho TW_{ikj}T_t$$
$$+ \gamma \boldsymbol{X}_{ikj} + \alpha \boldsymbol{W}_{kj} + \boldsymbol{\omega}_j + u_{ikjt}, \quad (5.1)$$

where *SW* stands for the average hours per week that students are shown the lecture videos in class, *TW* stands for the average hours per week that the local math teacher devoted to watching the lecture videos during lesson preparation, and all the rest of the variables are as defined in equation (4.1).

Table 11 presents the DD estimates for the effects of the DT program on student performance in all three subjects. In columns 1 and 2, we adopt the same specifications as in columns 1 and 2 of table 3 to examine if the estimated treatment effects differ with or without the covariates regarding student and teacher characteristics. The corresponding DD estimates in these two columns are very similar and statistically indistinguishable, indicating the treatment intensity has little correlation with the included covariates. The estimate on TW (weekly hours the local math teacher spent on watching the lecture videos) is 0.062, suggesting that each additional hour that the math teacher uses the recorded lectures improves their student's math scores by 0.062 standard deviations on average. Given the local teacher spends on average 6.42 hours doing so, the overall impact on students in the partial treatment group would be 0.398, which is comparable to our baseline estimate for the partial treatment effect of 0.343 (column 2 of table 3). The estimate for the effect of direct student per-hour exposure to the lecture videos is 0.429, which, at the average of 1.27 hours of weekly exposure time, translates into an increase of 0.545 standard deviations in math score for students in the full treatment group. Because students in the full treatment group also benefit from improved instruction quality due to teacher's exposure to the recorded lectures, the total impact of the DT on these students is an increase of math score by 0.943 standard deviations (evaluate as the average hours of student and teacher exposures to the lecture videos), which is quite similar to our baseline estimate of 0.978 for the full treatment effect.

Given the marginal benefit of direct exposure of students to the lecture videos is so much higher than that of indirect exposure via the local math teacher, why do teachers not choose to increase the usage of the recorded lectures in class? There are three possible explanations. The first and foremost reason is that the comparison of the two marginal effect estimates is not meaningful, because the teacher watched the lecture videos in their entirety while the students watched only parts of the recorded lectures selected by the teacher, presumably the parts that could benefit students the most. In fact, there is some evidence that teachers were very selective when it comes to what to show in class. According to our survey, the local math teachers reduced, rather than increased, the hours that they show the lecture videos in class because in part the lectures are tailored to the most capable students in an elite urban middle school and are difficult for much-less capable students in the remote schools to follow. There are simply fewer and fewer recorded lecture contents that are suitable for the students as they progress into higher grades. The second explanation relates to the improvement in the quality of local teachers. For local math teachers, the DT program works like a self-administered professional training program that improves the teacher's skills in all aspects of teaching activities, including lecturing. As our theoretical model predicts, the benefit of showing the lecture videos in class decreases as the quality or human capital of the local teachers increases. As a result, the in-class usage of lecture videos falls. The third reason for reduced in-class use of the lecture videos is that local teachers are under pressure to spend more time preparing students for the end-of-middle-school assessment exams. But this change in instructional emphasis is common across the treatment and control classes and across all subjects.

To examine whether the direct and indirect exposures are complements or substitutes, we introduce an interaction term between *TW* and *SW* as an additional independent variable. The estimates from this regression specification are reported in column 3. The estimated coefficient on the interaction term is not only very small but also statistically insignificant, suggesting that *TW* and *SW* are independent of each other. The estimates related to the treatment intensity are not affected at all in magnitude. The only noticeable change is that the standard error associated with the estimate for the effect of direct exposure is inflated by nearly 5 times due apparently to high level of correlation between the interaction term *TW**SW* and *SW*.

It should be noted that *SW* and *TW* are likely correlated with teacher, class and school characteristics. However, as long as these factors are time-invariant they are controlled for by the teacher professional rank variable, treatment status dummies, and school fixed effects which we include in the regressions in table 11. Since some schools have multiple classes assigned to treatment and control groups, class-specific factors remain omitted from the specification of columns 1 through 3. To address this issue, we add the class-specific average pre-treatment math score as a covariate to proxy for the omitted variable. As the estimates in column 4 show, the estimated effects of exposure (weekly per hour) to the DT program are very comparable to their counterparts in the previous columns. Endogeneity biases in the DD estimates, if exist, are not substantive enough to undercut our inferences regarding the effects of exposure of different levels.

As falsification tests, we also implement the treatment intensity regression models of columns 2 and 3 using the Chinese and English test scores, respectively. The estimates are reported in columns 5 through 8. Consistent with the findings reported in table 3, the DT program, which is designed to improve student performance in math, has absolutely no statistically significant effect on student performance in either Chinese or English.

6. Concluding remarks

We provide an evaluation of the DT program as a means to address the cross-region inequality in educational resources and outcomes. While the DT program shares many common features with other CAI programs that have been the subject of research in the literature, the unique research design of this study allows us to gain new insights into the mechanisms of such programs in affecting academic outcomes and therefore make two major contributions to the literature. First, we are the first to recognize the role of the local teacher in deciding which parts and how much of the program provided resources (mainly the lecture videos) to use as a part of in-class lecture. This is important because a typical CAI program is not highly customized for the targeted students. Indiscriminate exposure of students to the program would be inefficient and contrary to the pedagogy of teaching at the right level. The math teachers in our sample chose to show, on average, less than 40% of the lecture videos in class. This may help explain why the impact of the DT program is larger than those reported by other studies. Second, we are the first to view a CAI program also as a self-administered and topic-specific teacher training program. It affords the local teacher the opportunity to enrich their knowledge on the subject matter and improve their teaching skills through learning by observing, which in turn help improve student performance.

Using data that we collected from nine middle schools that participated in the math-focused DT program and a difference-in-differences model, we find that the DT program has a positive and statistically significant effect on student performance in math and the improvement is attributable to direct exposure of students to high-quality lecturing by the remote teacher as well as increases in the quality of instruction of the local teacher. Our estimates show that on average the DT program increased math test scores by 0.978 standard deviations over the three-year middle school education, of which 0.343 standard deviations are attributable to improved instruction on the part of the local teacher. We also find that the positive impact of the DT program is cumulative and largely independent of student and teacher characteristics, and robust to a plethora of tests and alternative model specifications. Overall, our findings suggest that the DT program is an effective means to improve education outcomes of students in underserved areas and hence help close cross-region gaps in education.

It is worth noting that the implementation of the DT program entails little extra cost. The computer equipment and internet infrastructure are largely available in remote and rural schools in China, as noted in Bianchi et al. (2022). Since the videos are recorded as the lectures are normally delivered in an elite urban school without customization or editing, the production cost is minimal. The cost of distribution over the internet is also negligible. Weighing the large benefits of the DT program that we have identified against the small costs of implementation, we conclude that the DT program is not only effective in improving student performance but doing so at low cost.

It is also important to note that the implementation of the DT program at the school level depends crucially on the support of local teachers, because the use of new resources requires additional time and efforts. Therefore, policies that increase the incentive for the local teachers to make the best use of online resources will be vital for the success and sustainability of the program.

References

Acemoglu, D., Laibson, D., List, J. (2014). Equalizing Superstars: The Internet and the Democratization of Education. American Economic Review: Papers & Proceedings, 104(5): 523-527.

Alpert, W., Couch, K., Harmon, O. (2016). A Randomized Assessment of Online Learning. American Economic Review, 106(5): 378-382.

Angrist, J., Lavy, V. (2002). New Evidence on Classroom Computers and Pupil Learning. Economic Journal, 112(482): 735-765.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M. (2016). Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. NBER Working Paper No. 22746.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., S. Khemani. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. American Economic Journal: Economic Policy, 2(1):1-30.

Banerjee, A., Cole, S., Duflo, E., Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. Quarterly Journal of Economics, 122(3):1235-1264.

Battaglia, M., Lebedinski, L. (2015). Equal Access to Education: An Evaluation of the Roma Teaching Assistant Program in Serbia. World Development, 76(C): 62-81.

Beg, S., Lucas, A., Halim, W., Saif, U. (2019). Beyond the Basics: Improving Post-Primary Content Delivery through Classroom Technology. NBER Working Paper No. 25704.

Bettinger, E., Fairlie, R., Kapuza, A., Kardanova, E., Loyalka, P., Zakharov, A. (2020). Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function. NBER Working Paper No. 26967.

Bianchi, N., Lu, Y., Song, H. (2022). The Effect of Computer-Assisted Learning on Students' Long-Term Development. Journal of Development Economics 158, 102919.

Borghesan, E., Vasey. G. (2021). The Marginal Returns to Distance Education: Evidence from Mexico's Telesecundarias. Working Paper.

Borzekowski, D. (2018). A Quasi-Experiment Examining the Impact of Educational Cartoons on Tanzanian Children. Journal of Applied Developmental Psychology, 54: 53-59.

Borzekowski, D., Henry, H. (2011). The Impact of Jalan Sesama on the Educational and Healthy Development of Indonesian Preschool Children: An Experimental Study. International Journal of Behavioral Development, 35(2): 169-179.

Borzekowski, D., Lando, A., Olsen, S., Giffen, L. (2019). The Impact of an Educational Media Intervention to Support Children's Early Learning in Rwanda. International Journal of Early Childhood, 51(1): 109-126.

Desai, S., Kulkarni, V. (2008). Changing Educational Inequalities in India in the Context of Affirmative Action. Demography, 45: 245-270.

Figlio, D., Rush, M., Yin, L. (2013). Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. Journal of Labor Economics, 31(4): 763-784.

Fleisher, B., Li, H., Zhao, M. (2010). Human Capital, Economic Growth, and Regional Inequality in China. Journal of Development Economics, 92: 215-231.

Fowler, R. (2003). The Massachusetts Signing Bonus Program for New Teachers: A Model of Teacher Preparation Worth Copying? Education Policy Analysis Archives, 11.

Glewwe, P., Kremer, M., Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. American Economic Journal: Applied Economics, 1:112–135.

Goolsbee, A., Guryan, J. (2006). The Impact of Internet Subsidies in Public Schools. Review of Economics and Statistics, 88(2): 336-347.

Harris, D., Sass, T. (2011). Teacher Training, Teacher Quality and Student Achievement. Journal of Public Economics, 95: 798-812.

Jacob, B., Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. Review of Economics and Statistics, 86(1): 226-244.

Johnston, J., Ksoll, C. (2017). Effectiveness of Interactive Satellite-Transmitted Instruction: Experimental Evidence from Ghanaian Primary Schools. CEPA Working Paper No. 17-08.

Kane, T., Rockoff, J., Staiger, D. (2008). What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City. Economics of Education Review, 27(6): 615-631.

Lai, F., Luo, R., Zhang, L., Huang, X., Rozelle, S. (2015). Does Computer-assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing. Economics of Education Review, 47: 34-48.

Lavy, V., Schlosser, A. (2005). Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits. Journal of Labor Economics, 23(4): 839-874.

Liu, Z. (2005). Institution and Inequality: the *Hukou* System in China. Journal of Comparative Economics, 33: 133-157.

Loeb, S., Page, M. (2000). Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market Opportunities and Non-Pecuniary Variation. Review of Economics and Statistics, 82(3): 393-408.

Loyalka, P., Popova, A., Li, G., Shi, Z. (2019). Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program. American Economic Journal: Applied Economics, 11(3): 128-154.

Ma, Y., Fairlie, R., Loyalka, P., Rozelle, S. (2020). Isolating the “Tech” From Edtech: Experimental Evidence on Computer Assisted Learning in China. NBER Working Paper No. 26953.

Matsudaira, J. (2008). Mandatory Summer School and Student Achievement. Journal of Econometrics, 142(2): 829–850.

Master, B. (2014). Staffing for Success: Linking Teacher Evaluation and School Personnel Management in Practice. Educational Evaluation and Policy Analysis, 36: 207-227.

Mcpherson, M., Bacow, L. (2015). Online Higher Education: Beyond the Hype Cycle. Journal of Economic Perspectives, 29(4): 135-153.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., Qiao, Y., Boswell, M., Rozelle, S. (2014). Integrating Computer-Assisted Learning into a Regular Curriculum: Evidence from a Randomized Experiment in Rural Schools in Shaanxi. Journal of Development Effectiveness 6(3): 300-323.

Mo, D., Bai, Y., Shi, Y., Abbey, C., Zhang, L., Rozelle, S., Loyalka, P. (2020). Institutions, Implementation, and Program Effectiveness: Evidence from a Randomized Evaluation of Computer-Assisted Learning in Rural China. Journal of Development Economics 146, 102487.

Muralidharan, K., Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. Journal of Political Economy, 119(1): 39-77.

Naik, G., Chitre, C., Bhalla, M., Rajan, J. (2020). Impact of Use of Technology on Student Learning Outcomes: Evidence from a Large-Scale Experiment in India. World Development, Volume 127: Article 104736.

Näslund-Hadley, E., Parker, S., Hernandez-Agramonte, J. (2014). Fostering Early Math Comprehension: Experimental Evidence from Paraguay. Global Education Review, 1(4):135-54.

Navarro-Sola, L. (2021). Secondary Schools with Televised Lessons: The Labor Market Returns of the Mexican Telesecundaria. Working Paper.

Prince, C. (2002). Higher Pay in Hard-to-Staff Schools: The Case for Financial Incentives. Arlington, VA: American Association of School Administrators.

Taylor, E., Tyler, J. (2012). The Effect of Evaluation on Teacher Performance. American Economic Review, 102(7): 3628-3651.

Wennersten, M., Quraishy, Z., Velamuri, M. (2015). Improving Student Learning via Mobile Phone Video Content: Evidence from the Bridge IT India Project. International Review of Education, 61(4), 503–528.

Zhang, Y. (2006). Urban-Rural Literacy Gaps in Sub-Saharan Africa: The Roles of Socioeconomic Status and School Quality. Comparative Education Review, 50(4): 581-602.

Table 1— Summary statistics of the main sample

Variables	N	Mean	SD	Min	Max
A. Program characteristics					
Number of students	1887				
F (full treatment class indicator=1 or 0)	1887	0.304	0.460	0	1
P (partial treatment class indicator=1 or 0)	1887	0.288	0.453	0	1
C (control class indicator=1 or 0)	1887	0.409	0.492	0	1
T (treatment period indicator=1 or 0)	1887	0.500	0.500	0	1
Number of teachers of treatment classes	12				
TW (average hours per week the teacher spent on watching lecture videos during lesson preparation)	12	6.42	0.558	5.25	7.25
SW (average hours per week the full treatment classes spent on watching lecture videos in class)	12	1.27	0.149	1.03	1.50
B. Student characteristics					
Gender (male=1, female=0)	1887	0.495	0.500	0	1
Rural (from rural areas=1, otherwise=0)	1887	0.580	0.494	0	1
Father education (years of schooling)	1887	8.396	3.572	0	16
Mother education (years of schooling)	1887	7.096	4.014	0	16
C. Teacher characteristics					
Number of teachers of treatment and control classes	25				
Advanced professional title	25	0.311	0.463	0	1
Intermediate professional title	25	0.547	0.498	0	1
Elementary professional title	25	0.142	0.349	0	1

Notes: The sample consists of 1887 students and 25 math teachers from 9 middle schools in underdeveloped regions in China.

Table 2—Summary statistics of test scores and cross-group differences

Variables	Average pre-treatment test scores			Average post-treatment test scores		
	Math	Chinese	English	Math	Chinese	English
F (Full Treatment)	-0.0262	0.0051	0.0109	0.556	0.0727	0.0448
P (Partial Treatment)	0.0315	0.0370	0.0358	-0.0213	0.0148	0.0155
C (Control)	-0.0027	-0.0299	-0.0312	-0.398	-0.0644	-0.0442
Cross-group difference						
F-C	-0.0198	0.0467	0.0511	0.977***	0.144	0.109
	(0.213)	(0.120)	(0.171)	(0.135)	(0.129)	(0.124)
P-C	0.0452	0.0772	0.0772	0.405***	0.0647	0.0365
	(0.117)	(0.0961)	(0.112)	(0.108)	(0.0950)	(0.107)
F-P	-0.108	-0.0531	-0.0425	0.566***	0.0563	0.0182
	(0.154)	(0.0933)	(0.136)	(0.145)	(0.131)	(0.142)

Notes: 1. All test scores are standardized by school and period; 2. Cross-group differences are estimated from school fixed-effects regression of the test score in a subject on treatment indicators; 3. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 3—The baseline results and placebo tests

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math	Math	Chinese	Chinese	English	English
F*T	0.978***	0.978***	0.102	0.102	0.0191	0.0191
	(0.144)	(0.144)	(0.143)	(0.143)	(0.182)	(0.182)
P*T	0.343**	0.343**	0.0124	0.0124	-0.0520	-0.0520
	(0.136)	(0.136)	(0.105)	(0.105)	(0.136)	(0.136)
F	-0.0114	-0.100	0.0377	-0.0284	0.0436	-0.101
	(0.205)	(0.189)	(0.118)	(0.105)	(0.166)	(0.139)
P	0.0443	-0.00452	0.0693	0.0298	0.0683	-0.0561
	(0.119)	(0.113)	(0.0954)	(0.0828)	(0.113)	(0.0982)
T	-0.396***	-0.396***	-0.0346	-0.0346	0.00882	0.00882
	(0.0902)	(0.0903)	(0.0870)	(0.0871)	(0.114)	(0.114)
Male		-0.0303		-0.348***		-0.335***
		(0.0415)		(0.0505)		(0.0471)
Rural *hukou*		0.0230		-0.0457		-0.0936
		(0.0627)		(0.0503)		(0.0629)
Father education		0.0351***		0.0286***		0.0293***
		(0.00679)		(0.00685)		(0.00774)
Mother education		0.0301***		0.0324***		0.0386***
		(0.00714)		(0.00649)		(0.00632)
Advanced title		0.0893		0.0279		0.517***
		(0.127)		(0.129)		(0.106)
Intermediate title		0.345*		0.0734		0.606***
		(0.172)		(0.152)		(0.115)
Constant	-0.0404	-0.329**	-0.0391	-0.0730	-0.0375	-0.915***
	(0.104)	(0.131)	(0.109)	(0.142)	(0.173)	(0.191)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,774	3,774	3,774	3,774	3,256	3,256
R-squared	0.081	0.128	0.002	0.068	0.001	0.088

Notes: 1. The dependent variable is pre- and post-treatment test scores in the subject; 2. The sample size reduction in columns 5 and 6 is due to missing pre-treatment English test scores for students from one of our sample schools. 3. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 4—Balance tests

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Pre-treatment math score	Pre-treatment Chinese score	Pre-treatment English score	Teacher with advanced rank	Teacher with intermediate rank
F	-0.0236	0.0365	0.0437	-0.123	0.120
	(0.206)	(0.119)	(0.170)	(0.160)	(0.124)
P	0.0345	0.0685	0.0686	-0.0852	0.0810
	(0.121)	(0.0959)	(0.114)	(0.163)	(0.127)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	1,887	1,887	1,628	1,887	1,887
R-squared	0.001	0.001	0.001	0.187	0.532

	(6)	(7)	(8)	(9)	(10)
VARIABLES	Pre-treatment math score	Male	Rural	Father education	Mother education
F	-0.139	-0.0208	-0.00727	1.034***	0.725**
	(0.190)	(0.0297)	(0.0214)	(0.289)	(0.291)
P	-0.0281	-0.0374	0.00641	0.659**	0.133
	(0.117)	(0.0288)	(0.0250)	(0.258)	(0.268)
Advanced title	0.0781				
	(0.159)				
Intermediate title	0.341				
	(0.227)				
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	1,887	1,887	1,887	1,887	1,887
R-squared	0.088	0.019	0.556	0.112	0.167

Notes: 1. None of the model specifications, except for column 6, includes any covariates. 2. Column 6 includes student and teacher characteristics as covariates to test if the pre-treatment math score is correlated with teacher characteristics. 3. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 5—Robustness checks: imbalance of parental education and non-random assignment issues

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Math	Math	Math	Math	Math	Math	Math
F*T	0.995***	1.023***	1.039***	1.164***	1.113***	0.960***	1.088**
	(0.152)	(0.154)	(0.162)	(0.197)	(0.173)	(0.221)	(0.398)
P*T	0.352**	0.387**	0.397***	0.518**	0.368**	0.333**	0.329*
	(0.136)	(0.147)	(0.147)	(0.207)	(0.173)	(0.121)	(0.169)
F	-0.114	-0.123	-0.138	-0.105	-0.298	-0.214	-0.674
	(0.195)	(0.194)	(0.199)	(0.189)	(0.214)	(0.354)	(0.412)
P	-0.0104	-0.0254	-0.0337	-0.00469	-0.0502	-0.0345	-0.138
	(0.114)	(0.128)	(0.130)	(0.114)	(0.116)	(0.214)	(0.199)
T	-0.396***	-0.440***	-0.440***	-0.396***	-0.428***	-0.377***	-0.352**
	(0.0903)	(0.106)	(0.106)	(0.0903)	(0.109)	(0.113)	(0.156)
Student chara.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher chara.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,648	3,428	3,302	3,774	2,832	1,532	914
R-squared	0.119	0.113	0.104	0.130	0.122	0.138	0.140

Notes: 1. Column 1 is based on the subsample excluding from the treatment group students whose parents received post-secondary education. 2. Column 2 is based on the subsample excluding from the control group students whose parents received no formal schooling. 3. Column 3 is based on the subsample excluding from the control group students whose parents received no formal schooling and from the treatment group students whose parents received post-secondary education. 4. Column 4 is based on the baseline sample (full sample) and includes interaction terms of father's and mother's schooling with F*T and P*T. None of these interaction terms obtains a significant coefficient estimate (not reported in the table). 5. Columns 5 through 7 are based respectively on the subsample of schools that randomly made the initial class assignment of students, subsample of schools that randomly assigned math teacher to the treatment class, and subsample of schools that randomly assigned both class and math teacher the treatment status. 6. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 6—Accounting for pre-treatment trends and triple difference in differences estimates

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math (with Chinese score as a control variable)	Math (with English score as a control variable)	Math (with average Chinese and English scores as a control variable)	Math minus Chinese score	Math minus English score	Math minus average score of Chinese and English
F*T	0.934***	0.977***	0.967***	0.876***	0.967***	0.951***
	(0.138)	(0.144)	(0.147)	(0.170)	(0.173)	(0.165)
P*T	0.337**	0.380***	0.367***	0.330*	0.406**	0.377***
	(0.145)	(0.124)	(0.124)	(0.175)	(0.150)	(0.136)
F	-0.0882	-0.122	-0.116	-0.0721	-0.0717	-0.0682
	(0.159)	(0.173)	(0.173)	(0.135)	(0.142)	(0.138)
P	-0.0174	-0.0545	-0.0476	-0.0343	-0.0265	-0.0183
	(0.110)	(0.107)	(0.114)	(0.123)	(0.103)	(0.108)
T	-0.381***	-0.397***	-0.390***	-0.361***	-0.401***	-0.388***
	(0.0907)	(0.0902)	(0.0925)	(0.112)	(0.112)	(0.105)
Student chara.	Yes	Yes	Yes	Yes	Yes	Yes
Teacher chara.	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,774	3,256	3,256	3,774	3,256	3,256
R-squared	0.302	0.375	0.381	0.083	0.105	0.115

Note: 1. Columns 1 through 3 repeat column 1 of table 3 with Chinese test score, English test score, and both Chinese and English test scores, respectively, as controls for group-specific common trends in academic performance. The estimates for these additional controls (not reported in the table) are positive and statistically significant at the one percent level. 2. Columns 4 through 6 repeat column 1 of table 3 with the dependent variable defined as the differences between math and Chinses test scores, math and English test scores, and math and the average of Chinese and math scores, respectively, assuming the trend of Chinese test score or English test score or the average of the two represents the counterfactual trend of math test score in the absent of the DT program. 3. When English test score is used the sample size reduces to 3256 because of missing pre-treatment English test score for students from one of our sample schools. 4. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 7—Additional robustness checks: difference in differences estimates with math as treated subject and non-math subject as control

VARIABLES	(1) The control subject is Chinese	(2) The control subject is English	(3) The control subject is average of Chinese and English
Panel A: Full treatment group			
Math *T	0.515***	0.566***	0.563***
	(0.132)	(0.138)	(0.133)
Math (the treated subject)	-0.0314	-0.0415	-0.0382
	(0.0981)	(0.112)	(0.110)
T	0.0675	0.0280	0.0310
	(0.117)	(0.147)	(0.132)
Observations	2,292	1,936	1,936
R-squared	0.155	0.186	0.196
Panel B: Partial treatment group			
Math*T	-0.0307	0.00557	-0.0112
	(0.139)	(0.104)	(0.0901)
Math (the treated subject)	-0.00545	-0.0132	-0.00576
	(0.0670)	(0.0652)	(0.0683)
T	-0.0222	-0.0432	-0.0265
	(0.0613)	(0.0769)	(0.0495)
Observations	2,172	1,816	1,816
R-squared	0.057	0.076	0.079
Panel C: Control group			
Math*T	-0.361***	-0.401***	-0.388***
	(0.114)	(0.114)	(0.107)
Math (the treated subject)	0.0272	0.0378	0.0306
	(0.0967)	(0.0692)	(0.0725)
T	-0.0346	0.00882	-0.00431
	(0.0884)	(0.116)	(0.0925)
Observations	3,084	2,760	2,760
R-squared	0.102	0.114	0.122

Notes: 1. All regressions include school fixed effects and all the covariates contained in the baseline regression of column 2 of table 3. 2. In all regressions, math is the treated subject with the treatment indicator Math=1 for math and Math=0 for other subjects. The control subject is Chinese in column 1, English in column 2, and average of Chinese and English (test scores) in column 3. 3. The dependent variable is the pre-treatment and post-treatment math scores for the treated subject and the corresponding Chinese (English, or average of Chinese and English) test scores for the control subject. 4. T=0 for pre-treatment period and T=1 for post-treatment period. 5. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 8—Difference in differences estimates by semester

VARIABLES	(1) 1st semester	(2) 2nd semester	(3) 3rd semester	(4) 4th semester	(5) 5th semester	(6) 6th semester
F*T	0.426***	0.513***	0.536**	0.495**	0.905***	0.978***
	(0.147)	(0.149)	(0.199)	(0.185)	(0.193)	(0.144)
P*T	0.103	0.176*	0.147	0.139	0.326**	0.343**
	(0.122)	(0.0986)	(0.126)	(0.108)	(0.157)	(0.136)
F	-0.125	-0.116	-0.107	-0.113	-0.0947	-0.100
	(0.187)	(0.187)	(0.191)	(0.187)	(0.196)	(0.189)
P	-0.0192	-0.0155	-0.00806	-0.0135	-0.000388	-0.00452
	(0.113)	(0.117)	(0.114)	(0.118)	(0.112)	(0.113)
T	-0.159	-0.206**	-0.205*	-0.190**	-0.369***	-0.396***
	(0.0967)	(0.0842)	(0.111)	(0.0908)	(0.112)	(0.0903)
Student chara.	Yes	Yes	Yes	Yes	Yes	Yes
Teacher chara.	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,774	3,774	3,774	3,774	3,774	3,774
R-squared	0.092	0.086	0.091	0.083	0.122	0.128

Notes: 1. The dependent variable is math test score from the end-of-semester final exam throughout the middle school education and the pre-treatment exam. 2. The final exam of the 6th semester is the standard end-of-middle-school exam. Therefore, column 6 is identical to the baseline results reported in column 2 of table 3. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 9--Pooled six semesters plus the pre-treatment scores

	(1)	(2)	(3)
VARIABLES	Math	Chinese	English
F* Time	0.138***	-0.00225	-0.0110
	(0.0199)	(0.0220)	(0.0231)
P*Time	0.0513***	-0.0109	-0.0166
	(0.0178)	(0.0159)	(0.0168)
Time	-0.0568***	0.00382	0.00791
	(0.0113)	(0.0135)	(0.0148)
F	0.0474	0.139	0.0818
	(0.141)	(0.101)	(0.149)
P	0.0249	0.111	0.00994
	(0.0992)	(0.0745)	(0.113)
Student chara.	Yes	Yes	Yes
Teacher chara.	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes
Observations	13,209	13,209	11,396
R-squared	0.109	0.083	0.091

Notes: 1. The dependent variable for each subject is the test score from the end-of-semester final exams and the pre-treatment test. 2. Time takes the value from 0 (for pre-treatment) to 6 (the last semester of middle school education). 3. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 10—Effect heterogeneity

	F*T	P*T	R^2	N
Panel A: By student gender				
Male	0.990***	0.301*	0.135	1868
	(0.163)	(0.161)		
Female	0.966***	0.381***	0.133	1906
	(0.143)	(0.134)		
Panel B: By student *hukou* status				
Rural	1.149***	0390**	0.153	2190
	(0.201)	(0.157)		
Urban	0.779***	0.279*	0.128	1584
	(0.099)	(0.153)		
Panel C: By father's education				
Less than middle school	1.016***	0.047	0.128	1188
	(0.207)	(0.185)		
Middle school or higher	1.000***	0.494***	0.118	2586
	(0.153)	(0.140)		
Panel D: By mother's education				
Less than middle school	0.973***	0.287*	0.127	1798
	(0.179)	(0.172)		
Middle school or higher	0.992***	0.390**	0.119	1976
	(0.156)	(0.150)		
Panel E: By teacher professional rank				
Advanced	1.100***	0.139	0.186	1172
	(0.186)	(0.106)		
Intermediate	0.958***	0.486**	0.132	2066
	(0.230)	(0.182)		
Elementary	0.891**	0.242	0.130	536
	(0.241)	(0.536)		
Panel F: By pre-treatment math score				
Lower 50%	1.069***	0.330**	0.198	1888
	(0.147)	(0.161)		
Upper 50%	0.923***	0.419**	0.247	1886
	(0.142)	(0.151)		

Notes: 1. All the regressions have the same specification as the baseline model of column 2 of table 3, except that each excludes the covariate that the heterogeneity test is designed for. 2. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.

Table 11—The effects of treatment intensity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Math	Math	Math	Math	Chinese	Chinese	English	English
TW*T	0.062***	0.063***	0.063***	0.069***	0.010	0.009	0.005	0.006
	(0.022)	(0.022)	(0.023)	(0.023)	(0.017)	(0.017)	(0.022)	(0.022)
SW*T	0.429***	0.414***	0.374	0.412***	0.016	-0.081	-0.059	0.231
	(0.132)	(0.130)	(0.627)	(0.122)	(0.123)	(0.993)	(0.150)	(1.166)
TW*SW*T			0.006			0.015		-0.047
			(0.102)			(0.155)		(0.185)
CAPTM				0.690***				
				(0.059)				
F	0.010	-0.075	-0.075	-0.041	-0.019	-0.019	-0.070	-0.072
	(0.208)	(0.191)	(0.191)	(0.057)	(0.103)	(0.103)	(0.137)	(0.133)
P	0.017	-0.037	-0.036	-0.057	0.005	0.006	-0.099	-0.104
	(0.121)	(0.116)	(0.117)	(0.042)	(0.085)	(0.086)	(0.103)	(0.102)
T	-0.399***	-0.400***	-0.400***	-0.419***	-0.043	-0.042	0.003	0.001
	(0.090)	(0.091)	(0.091)	(0.091)	(0.086)	(0.086)	(0.113)	(0.113)
Male		-0.031	-0.031	-0.008	-0.348***	-0.348***	-0.335***	-0.335***
		(0.042)	(0.042)	(0.040)	(0.050)	(0.050)	(0.047)	(0.047)
Rural hukou		0.022	0.022	0.068	-0.046	-0.046	-0.094	-0.093
		(0.063)	(0.062)	(0.053)	(0.050)	(0.049)	(0.063)	(0.063)
Father educ		0.035***	0.035***	0.030***	0.029***	0.029***	0.029***	0.029***
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
Mother educ		0.030***	0.030***	0.024***	0.032***	0.032***	0.039***	0.039***
		(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)
Advanced		0.094	0.094	0.016	0.028	0.028	0.516***	0.525***
		(0.123)	(0.123)	(0.068)	(0.129)	(0.128)	(0.106)	(0.105)
Intermediate		0.341*	0.341*	0.069	0.073	0.074	0.606***	0.615***
		(0.170)	(0.170)	(0.078)	(0.152)	(0.151)	(0.115)	(0.115)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,774	3,774	3,774	3,774	3,774	3,774	3,515	3,515
R-squared	0.078	0.125	0.125	0.207	0.067	0.067	0.088	0.088

Notes: 1. The dependent variable is the post-treatment and pre-treatment math test score in columns 1 through 4, Chinese test score in columns 5 and 6, and English test score in columns 7 and 8. 2. TW is average hours per week the math teacher spent on watching the lecture videos, and SW is average hours per week the students in the full treatment group spent on watching the lecture videos in class. 3. Column 4 includes class average pre-treatment math (CAPTM) test score to control for possible correlations between the average initial class math performance and TW as well as SW. 4. Robust standard errors in parentheses are clustered at the class level, and *** p<0.01, ** p<0.05, * p<0.1.