Routledge
Taylor & Francis Group

Journal of Development Effectiveness

ISSN: 1943-9342 (Print) 1943-9407 (Online) Journal homepage: https://www.tandfonline.com/loi/rjde20

The impact of teacher professional development programs on student achievement in rural China: evidence from Shaanxi Province

Meichen Lu, Prashant Loyalka, Yaojiang Shi, Fang Chang, Chengfang Liu & Scott Rozelle

JOURNAL OF DEVELOPMENT EFFECTIVENESS
2019, VOL. 11, NO. 2, 105–131
https://doi.org/10.1080/19439342.2019.1624594

ARTICLE

The impact of teacher professional development programs on student achievement in rural China: evidence from Shaanxi Province

Meichen Lu[a,b], Prashant Loyalka[c,d], Yaojiang Shi[a], Fang Chang[a], Chengfang Liu[e] and Scott Rozelle[c]

[a]Center for Experimental Economics in Education, Shaanxi Normal University, Xi'an, Shaanxi, China; [b]Department of Agriculgural and Applied Economics, Univeristy of Georgia, Athens, GA, USA; [c]Rural Education Action Program, Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, USA; [d]Graduate School of Education, Stanford University, Stanford, CA, USA; [e]School of Advanced Agricultural Sciences, Peking University, Beijing, China

ABSTRACT
There is a significant gap in academic achievement between rural and urban students in China. Policymakers have sought to close this gap by improving the quality of teaching in rural areas through teacher professional development (PD) programs. However, there is limited evidence on the effectiveness of such programs. In this paper, we evaluate the impact of a PD program-National Teacher Training Program (NTTP) and find that the NTTP has no effect on math achievement. We also find that while the program has a positive effect on math teaching knowledge of teachers, it has no significant effect on teaching practices in the classroom. Taken together, these results indicate that teachers may have improved their knowledge for teaching from NTTP, but did not apply what they learned to improve teaching practices or student learning.

ARTICLE HISTORY
Received 2 December 2016
Accepted 11 May 2019

KEYWORDS
Teacher professional development; teacher knowledge; student achievement; impact evaluation; rural China

1. Introduction

Students in rural Chinese schools have less access to education than do their urban peers, and the education that they receive is of lower quality. Rural students have low levels of educational attainment and persistently drop out of school across all post-primary levels of schooling (Liu et al. 2009; Yi et al. 2012; Shi et al. 2015). Several empirical studies have found that, in rural junior high schools, the dropout rate ranges from 18 to 31 percent across a number of Chinese provinces (Yi et al. 2012; Mo et al. 2013; Loyalka et al. 2013; Li et al. 2015; Shi et al. 2015; Wang et al. 2015). In large cities, nearly 90 percent of students attend academic high school. In contrast, at most, only 20 to 30 percent of junior high graduates in rural areas attend academic high school (Liu et al. 2009; Li et al. 2017). Further, in terms of academic challenges, such as competitive school entrance exams, students from rural areas perform significantly worse than do urban students (Loyalka et al. 2014).

A number of factors could be related to rural students' underperforming their urban counterparts in educational outcomes. One potential reason for the achievement gap is the fundamental differences in family investment in, and parental care for, students in rural versus urban areas (World Bank 2001; Huang and Du 2007; Luo et al. 2009; Wang et al. 2009; Ye and Pang 2011). Research from developed countries has shown the quality of teaching to be another important

CONTACT Yaojiang Shi shiyaojiang7@gmail.com Center for Experimental Economics in Education, Shaanxi Normal University, Xi'an, Shaanxi 710119, China

factor that affects educational outcomes among both advantaged and disadvantaged students (Darling-Hammond 2000; Rivkin, Hanushek, and Kain 2005; UNESCO, 2006). Evidence shows that a one-standard deviation increase in teacher quality will lead to a 0.15 to 0.24 standard-deviation improvement in student math achievement (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007; Kane and Staiger 2008; Fryer 2017). In a survey of education polices and student achievement conducted in all 50 states, Darling-Hammond (2000) found that teacher preparation and certification in the United States significantly correlates with student academic performance. Such work highlights the importance of teacher quality in improving student academic performance.

There are fewer studies of teacher quality in developing countries, but those available confirm that differences in teacher quality can significantly affect student achievement. One such study found that, in Peru, teachers with a high aptitude in math increased student achievement on standardized math tests by approximately 0.09 standard deviations (Metzler and Woessmann 2012). In China, teachers of the highest professional rank more positively affect rural students' achievement than do teachers of a lower rank (Chu et al. 2015).

Although there may be a number of ways to improve the quality of teaching for rural students (e.g., improving incentives for teachers; Muralidharan and Sundararaman 2011; Muralidharan 2012; Loyalka et al. 2016), policymakers in developing countries have placed great stock in teacher professional development (PD) programs (Cobb 1999; Villegas-Reimers 2003; Vegas 2007). In theory, PD programs seek to help teachers gain subject-specific knowledge (Dadds 2001), use appropriate pedagogical practices (Darling-Hammond and McLaughlin 1995; Schifter, Russell, and Bastable 1999), develop positive attitudes toward teaching (Cobb 2000), and, ultimately, improve the learning of students (Villegas-Reimers 2003). The importance of PD is further supported by empirical evidence from developed countries. For example, in a review of experimental evaluations of PD programs in developed countries, Yoon et al. (2007) found that the academic achievement of students whose teachers participated in PD programs increased by 0.54 standard deviations compared to students whose teachers did not participate in PD. These positive findings lend credence to the efforts of policymakers from developing countries who believe that PD improves the quality of teaching in their rural areas.

In recent years, the Chinese government has invested heavily in teacher PD programs. In 2010, China's government launched the National Teacher Training Program (NTTP), the country's flagship teacher PD program (MOE and MOF, 2010).[1] Beyond improving teaching quality in general, one of the major goals of the NTTP is to improve the quality of teaching in rural regions and improve the learning of rural students (Ministry of Education and Ministry of Finance (MOE and MOF) 2010). Given the high level of investment in the NTTP and its ambitious goals, the program is currently one of the key national government initiatives for improving the human capital of rural students and the equity of educational outcomes between rural and urban students in China. To improve the PD of teachers, the Ministry of Education has prescribed that the training content should focus on ethics in teaching, subject-specific knowledge, and pedagogical practices in proportions of 10, 40, and 50 percent, respectively (MOE, 2012).

Although policymakers in China and other developing countries invest billions of dollars in teacher PD programs each year (e.g., Yan, 2013; Government of Chile 2003; Government of India 2013), there is only limited evidence regarding whether these programs are effective (Bruns and Luque 2014). At best, policymakers have obtained only subjective feedback from the teachers who have been trained and the trainers who run the programs (e.g., Zuo and Su 2012). Although researchers also have compared different PD programs, studied the potential of new teaching technologies, and assessed teacher learning and other outcomes within these programs, using quantitative approaches (Garet et al. 2001; Borko 2004; Olakulehin 2007; Overbaugh and Lu 2008; Owston et al. 2008), few studies have demonstrated a causal relationship between the evaluated programs and student outcomes. In particular, there have been few large-scale, rigorous evaluations of the effects of teacher PD programs on student achievement in developing countries (Table 1).

One of these exceptions that we know of is Yoshikawa et al.'s (2015) study, which assesses the impact of a pilot PD program for early childhood education teachers in 64 preschools in Chile. Yoshikawa et al. found that the PD program moderately affected emotional and instructional support as well as classroom organization. The study's results, however, do not indicate any impact of the program on the cognitive outcomes of students.

Further, to our knowledge, there have been few large-scale empirical evaluations of PD programs in China. The only example that we know of is a study by Zhang et al. (2013) that involved 123 English teachers from 70 migrant schools in Beijing. Zhang et al. found that a short-term PD program had no significant impact on either teacher performance or student English test scores. The PD program evaluated, however, was created by the research team rather than by the government. There has been no large-scale empirical evaluation of the NTTP in China. Only one recent study describes a current teacher PD program system in China (Liu, Liu, C, and Loyalka 2016) but failed to evaluate the impact of PD programs on student achievement.

The overall goal of this study is to examine the effect that the NTTP has on student achievement in rural China. To this end, we first examine the impact of the NTTP on the math achievement of rural students. Next, we seek to understand the causal mechanism through which the NTTP affects student achievement. To do so, we examine whether the NTTP was able to improve either the knowledge of math or the teaching practices of teachers. We also seek to develop an explanation of the effect (positive, negative, or none) that the NTTP might have on students.

The remainder of the paper is organized as follows. In Section 2, we provide the background of the teacher professional development programs in China. In Section 3, we describe the conceptual framework for how a teacher PD program might be expected to improve student achievement. In Section 4, we present the study's methodology, including the sampling process, the nature of the intervention, our data collection effort, and our analytical approach. Section 5 provides the results of the study, and Section 6 concludes.

2. Background: teacher professional development programs in China

In China, there is a considerable number of government-run teacher PD programs. The largest program, run by the national government, is the NTTP. Others are run by regional and local governments – provincial, prefectural, and county. There are five main areas of differences between the NTTP and non-NTTPs: sources of funding, goals and target populations, training format, opportunities to participate in the training, and total hours of training.

The first difference between the NTTP and non-NTTPs is the source of funding. The NTTP is the highest level of a teacher PD program in China and is the only program that is supervised and funded by the central government. Each province has a provincial NTTP office that is responsible for organizing the NTTP training under the instruction of the central government. The financial investment in the NTTP is larger than that for non-NTTPs, which are organized by different (lower) levels of government. For example, the provincial non-NTTPs are organized and funded by provincial governments, and the prefecture non-NTTPs are organized and funded by prefecture governments.

The second difference is related to the goals and target populations of the programs. One of the major goals of the NTTP is to improve the quality of teaching in rural regions and the level of learning of rural students (Ministry of Education and Ministry of Finance (MOE and MOF) 2010). Thus, the main target population of the NTTP training is rural teachers who teach in China's nine-year compulsory education system. Although non-NTTPs also pay attention to improving the teaching quality of rural regions, their main focus is to promote PD for all teachers in their local areas. Thus, the target population of non-NTTPs is not only rural teachers

Table 1. Summary of studies on the impact of teachers PD on student learning.

Study (Study design)	Countries	Data (Year)	PD Intervention	PD measurement	Student outcome measurement	Results
Yoshikawa et al. 2015 (RCT)	Chile	119 teachers and 1876 students in 64 preschools	PD workshops and in-classroom coaching	Classroom Assessment Scoring System (in-classroom observation), no specific subject	Scores of picture vocabulary subtest, Language; Socioemotional skills index	Effect size on emotional support in class is 0.81. But no significant effect on child outcomes
Zhang et al. 2013 (RCT)	China	8387 students and 87 teachers in70 migrant schools (2009)	3-week training in English knowledge	Teacher standardised test scores, English	Student standardised test scores, English	No significant effect on either teacher or student test scores
Heller et al. 2012 (RCT)	USA	over 270 elementary teachers and 7,000 students (2007–2009)	Electric circuit courses	Percent correct scores of Content Knowledge Tests, Science	Percent correct scores of Content Knowledge Tests, Science	Effect size on teacher content knowledge is close to 2.0 and on student achievement is close to 0.4 to 0.6, both significant at 1% level.
Garet et al. 2008 (RCT)	USA	270 teachers and 5530 students in 90 elementary schools (2005–2007)	Eight content-focused institute and seminar days, or in-school coaching	Standardized scores of teachers' knowledge about instruction, Reading; Scores of instructional practice observed by enumerators	Standardized Average reading scores, Reading	Effect size on teacher knowledge is 0.37 to 0.38, but no significant effect on student achievement
Jacob and Lefgren 2004 (QED)	USA	100,228 students in 461 elementary schools (1996–1999)	Special funding for teacher development to probation school	No PD measurement	Scores of Iowa Test of Basic Skills	No significant effect on student achievement
Saxe and Gearhart 2001 (QED)	USA	23 classrooms	Integrating Mathematics Assessment (IMA); or Collegial Support (SUPP)	No PD measurement	Scores of test of fractions knowledge, Mathematics	Only IMA had significant effect on student achievement.

in the nine-year compulsory education system but also teachers in urban areas and high school teachers.

The third difference between the NTTP and non-NTTPs is that they differ in training format. The NTTP has two forms: on-site training and online training. We focus on the on-site training program and note that most non-NTTPs are conducted through on-site training.

Fourth, the number of slots for participation in the NTTP is relatively scarce as compared to the number that are offered for the non-NTTPs. For example, in Shaanxi Province, during each year from 2011 to 2013, only 1.6% of teachers were able to attend an on-site NTTP. In contrast, 14% were able to participate in an online NTTP (Liu, Liu, C, and Loyalka 2016).

Finally, the NTTP offers more training hours per course than do non-NTTPs. Liu, Liu, C, and Loyalka (2016) show that the NTTP (online and on-site) provided the most cumulative training hours per teacher in 2013 (33 hours), whereas provincial non-NTTPs offered 17 hours of training and prefecture non-NTTPs offered 19 hours.

3. Conceptual framework

Based on a review of literature on teacher PD programs (Chapman, Chen, and Postiglione 2000; Cohen and Hill 2000; Garet et al. 2001; Fishman et al. 2003; Yoon et al. 2007), we present a conceptual framework of the causal chain in which a PD program (here, the NTTP) ultimately raises student achievement. This framework assumes that teachers who attend the PD program will learn the materials taught during the program and that the material is designed such that teachers' instruction approaches will improve.

Research has shown that, if a PD program is not designed well, it results in only limited impacts (Ball and Cohen 1999; Yoshikawa et al. 2015). Several studies have found that one-day workshops often make teacher PD intellectually superficial, disconnected from deep issues of curriculum and learning, fragmented, and noncumulative (Ball and Cohen 1999; Wilson and Berne 1999). In a study of a PD program in Chile, Yoshikawa et al. (2015) found that a lack of content-focused, developmentally specific curricula was one of the leading reasons that the training failed to improve children's outcomes.

A well-designed PD program should be sustained, intensive, and content-focused to have a positive and lasting impact on classroom instruction and teacher performance (Ball and Cohen 1999; Wilson and Berne 1999; Garet et al. 2001; Yoon et al. 2007). One study found that providing textbooks and other reading material alongside the training; linking participation to incentives, such as promotion or salary implications; and maintaining a specific subject focus in the training, among other factors, resulted in positive impacts on student learning (Evans and Popova 2016). Further, program implementers most commonly mention providing mentoring follow-up visits, engaging teachers for their opinions and ideas, and designing programs in response to local context as related to positive impacts on student learning.

Figure 1 presents our framework. We examine three aspects related to the successful improvement of teacher PD for student achievement outcomes (teachers, students, and principal/school) across two phases (during and after PD training). In this framework, teachers are the direct participants in PD and are responsible for applying what they learned from PD to teach their students. Principals and schools provide the environment or institutional supports for teachers to participate in PD and to change their teaching practice. In the first phase (during PD training), teachers who receive training spots must attend and fully participate in the PD program. The PD programs must include an explicit focus on the subject matter to help teachers enhance their understanding and instruction of the subjects that they teach (Borko 2004; Yoon et al. 2007). That is, teachers must improve their teaching knowledge and/or skills through the PD program.

Further, the PD program itself should not disrupt student learning. For example, if the PD program is held during the semester at a central location, and teachers must be away from school

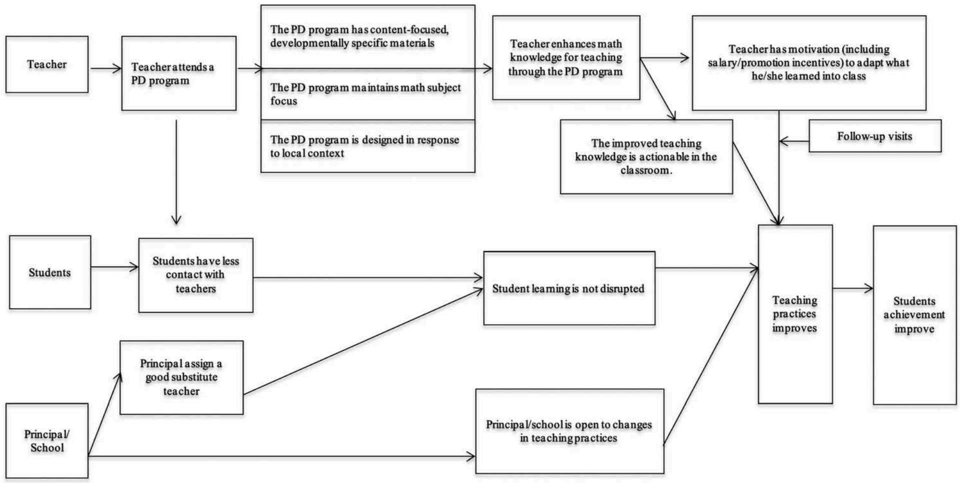

Figure 1. Framework for how the NTTP affects student achievement.

to receive the training, the full support of the principal or the school staff is needed to ensure that students are assigned a good substitute teacher.

In the second phase, after learning the material that was presented in the training program, teachers apply their learning in their classrooms. To incorporate their learning into their teaching practice, however, teachers must have the motivation, beliefs, and skills to overcome such barriers as limited time for preparation and instruction or limited materials and resources (Showers, Joyce, and Bennett 1987; Borko 2004). One of the effective ways to incentivize teachers to apply new teaching practice is performance pay, i.e., pay that is directly tied to student achievement scores (Muralidharan and Sundararaman 2011; Muralidharan 2012; Loyalka et al. 2016), which requires the schools to assess teachers based on student achievement.

The knowledge that teachers gain from the PD program must be suitable for their own students, i.e., is actionable in their own class. There are two conditions that determine whether the teaching knowledge is actionable. First, it is considered actionable when teachers adapt the new instructions from the PD program to the conditions of the classroom, enabling students to understand the subject matter well. This requires that the teaching knowledge provided by the PD program be based on a valid theory of action in regard to local classrooms (Hiebert and Grouws 2007; Yoon et al. 2007) and be aligned with the students' thinking. Second, new teaching knowledge is considered actionable when the school environment supports teachers in transforming their teaching practices. If teachers cannot adapt the new instructions to the current teaching resources, principals must be willing to provide new teaching resources. In this regard, the principal also should be open to changes in teaching practices. If all of these conditions are satisfied, teachers can successfully apply knowledge from PD programs into their teaching practices. Finally, the changes needed to ensure positive outcomes take time to develop and implement (King and Behrman 2009).

4. Method

4.1 Sampling

We conducted an evaluation of the NTTP for primary school math teachers in Shaanxi Province, located in northwest China. Its provincial GDP per capita is 46,928 yuan (approximately 7,640 US dollars); Shaanxi ranks 14th among the 31 mainland provinces of China in terms of GDP per capita (CNBS, 2014). Of the ten prefectures in Shaanxi Province, our study was conducted in two: Weinan, which ranks ninth in terms of GDP per capita among all Shaanxi prefectures, and Xianyang, which ranks fifth (Shaanxi statistical bureau 2014).

Teachers who participated in the evaluation were selected from a much larger group of teachers who had been chosen to receive NTTP training through a standard process across China. According to China's national teaching development policy, teachers are required to participate in a certain number of hours of PD training each year. The policy rules, however, contain no reference as to which type of program counts toward these hours. China's government offers non-NTTP PD programs at different levels – provincial, prefectural, and county (Liu, Liu, C, and Loyalka 2016). In the case of the (national) NTTP, the central office in Beijing decides which trainings will take place each year, and then each province's NTTP office allocates training slots to each prefecture. Each prefecture allocates the slots to each county. Finally, at the county education bureau level, officials choose the schools that are required to send teachers to that year's NTTP. The criteria used by county officials differ across regions, but, during interviews, we learned that schools were often given priority according to size and past rates of participation in the program. Specifically, the number of training slots allocated to a given school depends on the number of teachers in the school and past rates of participation (e.g., if the school sent a large number of math teachers to the NTTP in a previous year, the school might be allocated fewer slots in the NTTP for math teachers the next year). Although the number of slots assigned to each school in each year is different, on average, teachers of rural schools in the same county have roughly an equal opportunity to participate in NTTP for the subject they primarily teach. The exact registration process for individual teachers varies across schools.

We considered whether schools preferentially choose which teachers to send to the NTTP (e.g., giving the best teachers a slot in the training and ignoring the poorly-performing teachers). From our interviews with NTTP officials, local principals, and teachers, we believe that this is not the case. Although each school is responsible for collecting applications and submitting them to the county education bureau, every teacher reserves the right to apply to the program and to submit his or her application.

In fact, it is not the case that schools send only the best teachers to the NTTP. In some cases, schools prefer to keep strong teachers in the school because, in most schools, there are, for example, only one or two math teachers for each grade. If a teacher participates in the NTTP during the semester, it is often difficult for schools to find substitute teachers to take care of their students. Thus, some schools choose to send those teachers who teach less important (non-math) courses to the math NTTP. We excluded these teachers in our study by limiting our sample to teachers who taught math in one of our sample grade levels (3rd–6th grade). Moreover, because all sample schools have only one or two math teachers for each grade, it can be difficult to determine who is better to teach a certain grade.

To select our treatment sample for the evaluation, the research team first secured access to the list of all teachers who had been enrolled in the NTTP (for the year of our study). It should be noted that the research team was not involved in choosing the NTTP teachers; they already had been chosen by the time we launched our study. In the study prefectures, a total of 63 teachers were on the initial list of NTTP trainees.

The next step was to choose the treatment teachers from the overall list of trainees. To select a sample representative of the majority of schools in Xianyang and Weinan prefectures, we excluded teachers who came from schools that did not provide all six grades of primary school education. We further limited our sample to teachers who taught math in one of our sample grade levels (3rd–6th grade). After applying the exclusion criteria, we were left with 34 treatment teachers across 34 treatment schools.

The final step in choosing the evaluation sample was to determine which class and set of students were to be the beneficiaries of the NTTP. If a teacher in our sample taught math at more than one grade level, we randomly selected one grade level for inclusion in our study. If a teacher taught more than one math class in the same grade, we included all classes in that grade level in our sample.

We then selected teachers who did not participate in the NTTP to serve as a control group. Our control group comprised teachers selected in one of two ways. The first group of control teachers was selected from the same schools as the treatment teachers (*within-school control teachers*). The second group of control teachers was selected from different schools from the treatment teachers (*across-school control teachers*).

The within-school control teachers were selected using the following procedure: We called each treatment school to ask whether there were other math teachers who taught at the same grade level as our treatment teacher. If there was more than one other teacher, one was randomly selected as our control teacher. (If there was only one, that teacher was the control.) If there was not another math teacher in that grade level, then we did not select a within-school control teacher from that school. In total, we were able to sample 16 within-school control teachers. For each within-school control teacher who taught more than one math class at the same grade level as our treatment teacher, we randomly selected one of those classes for inclusion in our sample.

The across-school control teachers were selected according to the following protocol: First, we travelled to the Bureau of Education of each county where our treatment schools were located and collected information on all primary schools in that county. Specifically, we collected information on: (a) the distance between our treatment school and the other primary schools in that county; (b) the number of students in each school; and (c) the *tongkao* (standard exams organized by the Bureau of Education of each county) ranking of each school. We used these three criteria for several reasons.

First, the closer the control schools are to the treatment schools, the more similar they are in terms of local culture, school environment, administrative policy, student pool (i.e., students from similar backgrounds), and other factors at the school level that may be related to student academic performance. Second, the number of students represents school size, which the literature suggests is correlated with student performance (Kuziemko 2006). To rule out potential imbalances in this regard, we chose schools that had a similar number of students. Third, because the *tongkao* ranking is conducted at the county level, and students in each county were given the same standard tests, schools with a similar *tongkao* ranking reflect, to a certain extent, similar teaching and teacher quality. We used these three criteria to determine the overall nature of a school, which allowed us to choose a matched control school for each treatment school. Because schools were not randomly assigned to the treatment group – assignment to the treatment group was based on government rules – the best that we could do was to choose a matched control school for each treatment school by using these three criteria. In total, there were 34 pairs of treatment and control schools.[2]

After we selected the control schools, we used the following two-step procedure to select across-school control teachers: First, we contacted the principal of each school and asked for a list of math teachers who taught the same grade level as the treatment teacher from the matched school. Next, we randomly selected one teacher from that grade level for each control school. In total, we selected 34 across-school control teachers. If the teacher taught more than one math class in the same grade as our treatment teacher, we randomly selected one class for inclusion in our sample.

The sample at the time of the baseline survey (October 2014) consisted of 84 teachers (34 treatment teachers, 16 within-school control teachers, and 34 across-school control teachers) and 3,289 students. By the time of the endline survey in February 2015, however, there was some student attrition. Due to various reasons (dropouts, absences, death, and missing data), 223 students (71 students in the treatment group and 152 students in the control group) did not complete our endline survey. This means that the attrition rate was 7 percent. There was no attrition of teachers, however. Thus, in total, our final sample included 84 teachers and 3,066 students. In total, there were 1,141 students in the treatment schools and 1,922 students in the control schools.

To examine whether attrition of students affected our results, we regressed attrition status on the treatment variables. A comparison of the attrition rates between the treatment and control groups shows that treatment status did not affect the attrition rate (Table 2). Further, we checked for the balance on observable characteristics among the students and teachers who participated in both the baseline and endline surveys. As seen in Table 3, the characteristics of treatment and control students are fully balanced (Rows 1 to 8). In terms of teacher characteristics, Table 4 shows that one key outcome variable (standardized Math Knowledge for Teaching test scores; Row 1) and most of the control variables (including teacher gender, whether the teacher completed university, and whether the teacher has the highest rank) are balanced. We did find significant differences in terms of two variables: whether the teacher majored in math and his or her teaching experience (Table 4, Rows 4 and 6).[3] To account for the imbalance in these variables in our subsequent analyses, we controlled for them in our regressions. After adding control variables, we saw no significant change in the point estimates for each outcome, suggesting that the imbalance is not significant enough to affect internal validity.

4.2 Intervention

Our intervention was conducted at the teacher level. All treatment teachers participated in the NTTP after the baseline survey. The PD program in which they participated was organized by two top-ranked primary schools under the oversight of the Shaanxi provincial NTTP office. The research team did not have any input into the training. This was a training program run fully by the education system. The instructors in the program were blind to the study. During the two-week

Table 2. Comparisons of attrition between the treatment and control groups.

Dependent Variable: Attrition (1 = student left, 0 = student remained)	(1)	(2)
1. Treatment students (yes = 1; no = 0)	−0.02	−0.01
	(0.01)	(0.01)
School-pair fixed effect	Yes	Yes
Student Characteristics	NO	Yes
Observations	3289	3289
R-squared	0.001	0.004

Source: Authors' survey
*** p < 0.01, ** p < 0.05, * p < 0.1
Note: Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.

Table 3. Student characteristics at baseline.

	Treatment group	Control group	Difference: (1) – (2)	P-value
	(1)	(2)	(3)	(4)
Variable	Mean Sd	Mean Sd		
Student Characteristics				
1. Standardized math test score (SD)	−0.00 [1.00]	0.00 [1.00]	−0.01 (0.06)	0.91
4. Student age (year)	9.67 [1.41]	9.59 [1.39]	0.09 (0.08)	0.28
5. Female student (1 = yes, 0 = no)	0.48 [0.50]	0.47 [0.50]	0.01 (0.02)	0.63
2. Student is a left-behind child (1 = yes, 0 = no)	0.15 [0.36]	0.15 [0.36]	−0.00 (0.01)	0.94
3. Student lives at school (1 = yes, 0 = no)	0.07 [0.25]	0.08 [0.27]	−0.01 (0.02)	0.47
6. Mother completed junior high school (1 = yes, 0 = no)	0.41 [0.49]	0.41 [0.49]	0.00 (0.02)	0.92
7. Father completed junior high school (1 = yes, 0 = no)	0.41 [0.49]	0.43 [0.50]	−0.02 (0.02)	0.34
8. Teacher teaching practice index (reported by students, SD)	0.02 [0.44]	−0.01 [0.45]	0.03 (0.03)	0.43
Number of observations	1144	1922	3066	

Source: Authors' survey
*** p < 0.01, ** p < 0.05, * p < 0.1
Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.

Table 4. Teacher characteristics at baseline.

	Treatment group	Control group	Difference: (1) − (2)	P-value
	(1)	(2)	(3)	(4)
Variable	Mean Sd	Mean Sd		
Teacher Characteristics				
1. Teacher MKT test score (SD)	−0.03 [1.16]	−0.11 [1.07]	0.08 (0.26)	0.77
2. Female teacher (1 = yes, 0 = no)	0.85 [0.36]	0.82 [0.39]	0.03 (0.07)	0.62
3. Teacher completed university (1 = yes, 0 = no)	0.18 [0.39]	0.30 [0.46]	−0.12 (0.08)	0.13
4. Teacher majored in math (1 = yes, 0 = no)	0.15 [0.36]	0.32 [0.47]	−0.17* (1.00)	0.08
5. Teacher has highest rank (1 = yes, 0 = no)	0.15 [0.36]	0.24 [0.43]	−0.09 (0.09)	0.30
6. Teaching experience (1 = over 15 years, 0 = equal or below 15 years)	0.21 [0.41]	0.46 [0.50]	−0.25** (0.09)	0.01
7. Teacher who undergone provincial PD in the past year (1 = yes, 0 = no)	0.26 [0.45]	0.20 [0.40]	0.06 (0.09)	0.47
8. Teacher who undergone prefecture PD in the past year (1 = yes, 0 = no)	0.35 [0.48]	0.42 [0.50]	−0.07 (0.10)	0.49
Number of observations	34	50	84	

Source: Authors' survey.
*** p < 0.01, ** p < 0.05, * p < 0.1
Notes: Teacher MKT test score means teacher math knowledge for teaching test score.

program, all treatment teachers were released from regular teaching duties to receive training at a centralized location in the prefecture.

The curriculum of the treatment teachers' PD program followed the basic mandated framework of the NTTP. According to the teaching materials that we observed the treatment teachers using, there were sections of the training classes that were focused on ethics in teaching, subject-specific knowledge, and pedagogical practices (Ministry of Education (MOE) 2012). The instructors of the NTTP were top teachers from the highest-ranked primary schools in the prefecture. These teachers were selected as trainers due to their reputation for being knowledgeable about teaching math. Training sessions were conducted for six hours per day; sessions consisted primarily of lectures, with short intervals for lesson modeling, questions, and discussion.

The NTTP provided only the basic components of the in-program training and did not provide any in-classroom follow-up instruction. In other words, after the training, neither the NTTP instructors nor any other NTTP staff went to any of the schools of the treatment teachers. There was no attempt to observe whether teachers applied what they learned, to identify any weaknesses in the teaching practices of teachers, or to instruct teachers on how to address such weaknesses. The only post-training follow-up occurred through the establishment of an online group-chat for teachers, which was organized by the teachers themselves, not by the NTTP staff.

4.3 Data collection

We collected data based on student and teacher responses to a baseline survey (October 2014) and an endline survey (February 2015). The student baseline survey consisted of three blocks. In the first block, students provided basic demographic and family background information, including age, gender, whether he or she was a left-behind child,[4] boarding status, and whether his or her father or mother had completed at least junior high school.

In the second block, students completed a checklist of practices, which might have been influenced by participation in the PD program, that they observed in their teachers.[5] The checklist includes 13 items, consistent with those used in the Programme for International Student Assessment (PISA) survey (OECD 2014). Students responded to items that were presented with a 4-point Likert-type response of 'every lesson,' 'most lessons,' 'some lessons,' and 'never or hardly ever' in response to: How often do these things happen in your mathematics lessons: the teacher sets clear goals for our learning; teacher asks me or my classmates to present our thinking or reasoning at some length; the teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster; the teacher assigns projects that require at least one week to complete; the teacher tells me about how well I am doing in my mathematics class; the teacher asks questions to check whether we understood what was taught; the teacher has us work in small groups to come up with joint solutions to a problem or task; at the beginning of a lesson, the teacher presents a short summary of the previous lessons; the teacher asks us to help plan classroom activities or topics; the teacher gives me feedback on my strengths and weaknesses in mathematics; the teacher tells us what is expected of us when we get a test, quiz or assignment; the teacher tells us what we have to learn; the teacher tells me what I need to do to become better in mathematics.

Using student responses to this checklist, we created an index (teaching practice index) of a group of variables (the 13 items of teaching practice) according to the generalized least squares (GLS) weighting procedure described in Appendix A of Anderson (2008).[6] As the index is based not on teacher responses, but, rather, on students' observations of teachers, we believe that it is more likely to reflect teachers' actual practices.

In the third block of the student survey, students were given a 30-minute standardized math test. The math test items were drawn from the Chinese National Curriculum Framework (MOE, 2011). We prepared and administered the test ourselves to ensure that students and teachers could not prepare for the test. The enumeration team closely proctored the exams to strictly enforce time limits and minimize cheating. The scores were standardized by scaling them into *z*-scores by subtracting the mean and dividing by the standard deviation of the math score distribution of all students tested. These standardized scores are used as our key measure for assessing student baseline math achievement.

We administered the teacher baseline survey in two blocks. In the first block, enumerators collected information on the characteristics of teachers. Specifically, we solicited information on each teacher's gender, educational attainment, whether he or she majored in math, whether he or she had obtained the highest teaching rank, and his or her teaching experience.

In the second block of the teacher baseline survey, we administered the 40-minute Mathematical Knowledge for Teaching (MKT) test, designed by a research team at the University of Michigan (Study of Instructional Improvement and Learning Mathematics for Teaching Project; Hill, Schilling, and Ball 2004). According to the manual, the MKT test is designed not to test the specific math knowledge of teachers but, rather, to test the knowledge for teaching math to students. For example, the MKT test seeks to test teachers' knowledge of reasoning about mathematics and student thinking as well as how well teachers are able to solve mathematical problems that arise in the process of teaching. Thus, the MKT test was designed to be used in a number of contexts, no matter which curriculum teachers are expected to teach.

The versions of the test that we used comprise approximately 20 specific questions that ask teachers to explain how they would teach specific mathematics problems to students. We used two versions of the test: Version A (19 questions) and Version B (21 questions). They are inherently the same test and differ only in terms of expression and combination of questions. This test has been used in many studies (Delaney et al. 2008; Agodini et al., 2009; Bell et al. 2010; Copur-Gencturk and Lubienski 2013; Faulkner and Cain 2013; Hill, Rowan, and Ball 2005). The MKT test scores of teachers were standardized for ease of interpretation (Hill, Rowan, and Ball 2005; Daleny et al., 2008). A higher MKT score indicates that a teacher is more capable in the areas of teaching math.

In February 2015, we returned to all of the sample schools to conduct a follow-up survey. The procedures for conducting the endline survey were identical to those for the baseline survey with two exceptions. First, we did not ask students or teachers for their basic background information. Second, we gave Version B of the MKT to all teachers who had received Version A at baseline. We also gave Version A of the MKT to the teachers who had received Version B at baseline. We did this to avoid the issue of teachers' being able to recall the answers that they gave during the baseline survey.

4.4 Analytical approach

Here, we introduce the analytical approach we used to examine the impact of the NTTP on both student academic achievement and teachers' knowledge of math teaching practices. We assess the impacts, using ordinary least squares (OLS) regression with school-pair fixed effects.

4.4.1 OLS Model for examining the impact of the NTTP on student academic achievement

We use unadjusted and adjusted OLS regression analysis to estimate how student achievement in the treatment group changed relative to student achievement in the control group. The unadjusted model is:

$$Y_i = \alpha_0 + \alpha_1 T_i + \varphi_p + \varepsilon_i \tag{1}$$

Where Y_i represents the endline math achievement score of student i; T_i is a dummy variable that equals 1 if the math teacher of student i participated in the NTTP treatment, and 0 otherwise; and ε_i is a random error term. It also includes a school-pair fixed effect, φ_p. Note that, because there were 34 pairs of treatment schools and control schools in our sample, the matrix φ_p comprises 34 school-pair dummy variables.

To control for potential confounding effects of student and teacher characteristics, we also adjusted for additional covariates (X_i and Z_i). We call equation (2), below, our adjusted school-pair fixed effect model:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 Z_i + \varphi_p + \varepsilon_i \tag{2}$$

where X_i represents a vector of student characteristics at baseline, including *standardized baseline math test score, age, female student (1 = yes, 0 = no)*, student is a *left-behind child (1 = yes, 0 = no), student lives at school (1 = yes, 0 = no), mother completed junior high school (1 = yes, 0 = no)*, and *father completed junior high school (1 = yes, 0 = no)*. The additional term, Z_i, represents a vector of math teacher characteristics, including baseline *female teacher (1 = yes, 0 = no)*, whether the teacher *completed university (1 = yes, 0 = no)*, whether the teacher *majored in math (1 = yes, 0 = no)*, whether the teacher has the *highest rank (1 = yes, 0 = no)*, and *teaching experience (1 = over 15 years, 0 = equal to or below 15 years)*. See Tables 3 and 4 for the descriptive statistics of these variables. In addition, equation (2) also includes a school-pair fixed effect term φ_p.

To examine whether the NTTP had greater impacts on certain subgroups, we used a heterogeneous effects model to estimate treatment parameters. The heterogeneous effects model is essentially equation (2) with an additional interaction term between the NTTP treatment variable and a student background variable (measured using the baseline survey data). We chose several different types of student background variables to interact with the treatment variable, including *baseline student math test score, student age, female student*, and *student is a left-behind child*. In all regressions, we included a school-pair fixed effect.

4.4.2 OLS Model for examining the impact of the NTTP on math knowledge for teaching and teaching practices

The primary outcome variables for teachers were (a) math knowledge for teaching (MKT) test score and (b) the teaching practices index. As with the examination of the impacts of the NTTP on student outcomes, we first conducted unadjusted analyses, using the following model:

$$\mathrm{Y_j} = \eta_0 + \eta_1 \mathrm{T_j} + \varphi_p + \varepsilon_j \tag{3}$$

where Y_j represents the outcome variable of teacher j; T_j is a dummy variable that equals 1 if the teacher participated in the NTTP and 0 otherwise; and φ_p is a school-pair fixed effect.

We also conducted an adjusted analysis, which controls for teacher-level covariates (Z_j). The adjusted school-pair fixed effect model is as follows:

$$\mathrm{Y_j} = \lambda_0 + \lambda_1 \mathrm{T_j} + \lambda_2 Z_j + \varphi_p + \varepsilon_j \tag{4}$$

where Z_j represents a vector of math teacher characteristics, including *baseline MKT test score/baseline math teaching practice index*, teacher *gender*, whether the teacher *completed university*, whether the teacher *majored in math*, whether the teacher has the *highest rank*, and *teaching experience*.

5. Results

5.1 Effect of the NTTP on student academic achievement

Our analyses suggest that the NTTP has no impact on student academic achievement. In the unadjusted model, the estimate of the impact of the NTTP on student achievement was small in magnitude and not statistically significant (Table 5, Column 1, Row 1). After adding controls for student and teacher characteristics and school fixed effects, we found that the NTTP actually produces a small, negative impact on student academic achievement. The result was not significantly different from zero, however. Finally, results from the adjusted school-pair fixed effect model showed that the NTTP decreased the academic achievement of treatment students by 0.07 standard deviations (Table 5, Column 2, Row 1).[7] This result is statistically significant at the 10% level. In other words, at a minimum, the evidence suggests that teacher participation in the NTTP does not increase student achievement and may even harm student achievement. Given the government's sizeable investments in the NTTP, these results are concerning. Improving the achievement of rural students is one of the top goals of the nation's flagship teacher PD program.[8]

We also were interested in determining whether the NTTP has differential impacts on different subgroups of students. To test this, we examined heterogeneous effects by applying the treatment variable separately to each student background characteristic of interest (baseline student math test score, student age, student gender, and whether the student is a left-behind child). This analysis showed no significant heterogeneous impacts of the NTTP on student academic achievement for any of these variables (Table 6, Rows 2 to 5).

5.2 Unpacking the causal chain: effect of the NTTP on teacher math knowledge for teaching and teaching practices

Our analysis showed that the NTTP did not improve student academic achievement (and may even be harmful for student learning). One reason for this finding is that the NTTP was simply ineffective in teaching teachers new skills. To check for this possibility, we looked at the impact of the NTTP on teacher math knowledge for teaching. According to our unadjusted model, the NTTP increased the MKT test scores of treatment teachers by 0.55 standard deviations relative to control teachers (Table 7, Column 1, Row 1).[9] This result is significant at the 5% level. The results were consistent after controlling for school-pair fixed effects and teacher characteristics. We found treatment

Table 5. The effect of the NTTP on student academic achievement (Shaanxi Province, China).

Dependent variable: Endline student math score (SD)	(1) Unadjusted	(2) Adjusted
1.Teacher participates in teacher training (1 = yes, 0 = no)	0.00	−0.07*
	(0.04)	(0.04)
2. Baseline student math score (SD)		0.52***
		(0.02)
Student characteristics	NO	YES
Teacher characteristics	NO	YES
School-pair fixed effect	YES	YES
Constant	−0.00	1.63***
	(0.03)	(0.28)
Observations	3,066	3,066
R-squared	0.148	0.413

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Notes:
(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.
(b) The unadjusted model does not control for student characteristics or teacher characteristics, while the adjusted model does. The student characteristics include: student age, gender, whether the student is a left-behind child, whether the student lives at school, whether the student's mother completed junior high school, and whether the student's father completed junior high school.
(c) The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, and teaching experience, whether teacher undergone provincial PD in the past year, Teacher who undergone prefecture PD in the past year.
(d) The unadjusted model and the adjusted model both include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.
(e) Additionally, we ran another analysis that considers the two types of control teachers by adding a dummy variable that equals 1 if the teacher was a within-school control teacher and 0 otherwise. The results of this approach are substantively identical to the above estimates. They are reported in a supplemental appendix available online (https://reap.fsi.stanford.edu/publication/impact-teacher-professional-development-programs-student-achievement-rural-china).

teachers' average MKT test score to be 0.36 standard deviations higher than that of control teachers, which is significant at the 5% level (Table 7, Column 2, Row 1).

Although the NTTP failed to improve student achievement, our results tentatively suggest that the teachers may have learned more about how to teach math. This finding suggests that the null impact of the NTTP on student achievement cannot be attributed to a lack of improvement in the math knowledge for teaching. It is clear, however, that whatever knowledge teachers learned in the training was not translated into improved student achievement.

One possible explanation for the disconnection between teacher and student achievement is that, even if the NTTP improved teacher knowledge of teaching math, teachers may not have changed their teaching practices after completing the NTTP. Indeed, even though the Ministry of Education has prescribed that 50 percent of NTTP training material should focus on pedagogical practices, our results show that the effect of the NTTP on teaching practices is negligible. In the endline, the teaching practice index of the treatment teachers rises slightly, to 0.019 standard deviations, while the index of the control teachers falls slightly, to −0.011 standard deviations. Importantly, however, the difference between the teaching practice indices of the treatment and control teachers was not statistically significant. Although the unadjusted model shows that the teaching practice index of the average teacher in the treatment group is 0.03 standard deviations higher than that of the average teacher in the control group, this is not significant (Table 8, Column 1, Row 1).

The results from our adjusted school-pair fixed effect model also show that the NTTP has no significant impact on teaching practices (Table 8, Column 2, Row 1). In view of the possibility that student responses

Table 6. Heterogeneous effects of NTTP on student academic achievement (Shaanxi Province, China).

Dependent variable: Endline student math score (SD)	(1)	(2)	(3)	(4)
1.Teacher participates in teacher training (1 = yes, 0 = no)	−0.07**	0.22	−0.03	−0.08**
	(0.04)	(0.29)	(0.05)	(0.04)
2. Teacher participates in teacher training *Baseline student math test score	0.00			
	(0.04)			
3. Teacher participates in teacher training *Student age		−0.03		
		(0.03)		
4. Teacher participates in teacher training *Female student			−0.07	
			(0.07)	
5. Teacher participates in teacher training *Student is a left-behind child				0.10
				(0.09)
6. Baseline student math score (SD)	0.52***	0.52***	0.52***	0.52***
	(0.02)	(0.02)	(0.02)	(0.02)
7. Student age (years)	−0.16***	−0.15***	−0.16***	−0.16***
	(0.02)	(0.03)	(0.03)	(0.03)
8. Female student (1 = yes, 0 = no)	−0.04	−0.04	−0.01	−0.04
	(0.03)	(0.03)	(0.04)	(0.03)
9. Student is a left-behind child (1 = yes, 0 = no)	−0.01	−0.01	−0.01	−0.04
	(0.04)	(0.04)	(0.04)	(0.05)
Student characteristics	YES	YES	YES	YES
Teacher characteristics	YES	YES	YES	YES
School-pair fixed effect	YES	YES	YES	YES
Constant	1.61***	1.53***	1.61***	1.62***
	(0.24)	(0.25)	(0.28)	(0.28)
Observations	3,066	3,066	3,066	3,066
R-squared	0.413	0.413	0.413	0.413

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.
(b) The student characteristics include: whether the student lives at school, whether the student's mother completed junior high school, and whether the student's father completed junior high school.
(c) The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, and teaching experience, whether teacher undergone provincial PD in the past year, Teacher who undergone prefecture PD in the past year.
(d) The heterogeneous effects analyses include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.

are unreliable (and to carry out a supplemental approach), we conducted independent classroom observations in eight randomly selected classrooms (four control classrooms and four treatment classrooms)[10] but did not observe any difference in teaching styles between treatment and control teachers.

5.3 Discussion

Our results indicate that the NTTP fails to improve students' academic achievement. One possible reason is that teachers improved their teaching knowledge through NTTP but did not use the teaching knowledge they learned in the NTTP to improve their teaching practices.

We were interested in why teachers fail to translate their new teaching knowledge into their teaching practices. Although it was not possible to measure this quantitatively, there are two reasons why this may have happened. One possibility is that teachers may have lacked the ability or incentive to convert their increased knowledge into improved teaching practices and, ultimately, into student achievement. This might be due to the fact that, in most rural schools in China, workload and attendance are the main criteria that schools use to assess teachers, and student achievement is not directly linked to teacher performance pay (Shi et al. 2015). Thus, teachers may not have enough of an incentive to make the effort to change their teaching practices and to improve student achievement.

Table 7. The effects of the NTTP on the math knowledge for teaching (MKT) of teachers (Shaanxi Province, China).

Dependent variable: Endline teacher MKT test score (SD)	(1) Unadjusted	(2) Adjusted
1. Teacher participates in teacher training (1 = yes, 0 = no)	0.55**	0.36**
	(0.24)	(0.19)
2. Baseline teacher MKT test score (SD)		0.61***
		(0.14)
Teacher Characteristics	NO	YES
School-pair fixed effect	YES	YES
Constant	−0.27*	−0.07
	(0.16)	(0.48)
Observations	82	82
R-squared	0.465	0.735

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Notes:
(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.
(b) The unadjusted model does not control for teacher characteristics while the adjusted model does. The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, teaching experience whether teacher undergone provincial PD in the last year and whether teacher undergone prefecture PD in the last year.
(c) The unadjusted model and adjusted model both include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.
(d) Two teachers (one that participated in the NTTP and one that did not participate in the NTTP) were unable to complete the teacher MKT test in the endline survey. Thus we have 82 (instead of the 84 in our sample) teachers in this analysis.

Table 8. The effect of the NTTP on the teaching practices of teachers (Shaanxi Province, China).

Dependent variable: Endline teaching practice index (SD) (Report from Students)	(1) Unadjusted	(2) Adjusted
1. Teacher participates in teacher training (1 = yes, 0 = no)	0.03	−0.00
	(0.03)	(0.03)
2. Baseline teacher teaching practice index (SD)		0.32***
		(0.02)
Teacher Characteristics	NO	YES
School-pair fixed effect	YES	YES
Constant	−0.01	−0.07
	(0.02)	(0.05)
Observations	3066	3066
R-squared	0.068	0.170

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Notes:
(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.
(b) The unadjusted model does not control for teacher characteristics while the adjusted model does. The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, teaching experience whether teacher undergone provincial PD in the last year and whether teacher undergone prefecture PD in the last year.
(c) The unadjusted model and adjusted model both include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.

Another possibility is that, even if teachers have such motivation, the teaching knowledge that they learned from the NTTP may not have been helpful or appropriate for use in the classroom. In fact, teachers who were interviewed during pretests/piloting often complained that much of the material that they learned in NTTP sessions was not actionable. According to our interviews, some new teaching methods are impractical for rural students. For example, some teaching methods require students to collect information by themselves, which is impossible for many rural students who have no access to the Internet. Moreover, lack of technological support from the schools is another barrier for teachers to apply their learning to their teaching practices. Several interviewed teachers stated that some of the methods that they learned from the NTTP require the help of new technologies, such as multimedia graphing tools and computers, but that it is highly unlikely that every classroom in rural China will have such equipment. Therefore, it is difficult for the teachers to transfer what they learned from the NTTP into their teaching practices.

There also may have been offsetting effects. In particular, teachers' absence from their classrooms may have disrupted student learning. The NTTP evaluated in this study was conducted in urban areas far from the home schools of our sample teachers. As a result, teachers had to leave their classrooms for two weeks to participate in the training. In most rural schools, each grade has only one math teacher. Therefore, the absence of the trained teacher may have resulted in the substitution of teachers who are less familiar with the course curriculum and subject matter or less motivated to teach well in the treatment teacher's classroom. It is possible that this temporary disruption may have had a negative effect on student learning that offset or even exceeded any positive effects of the NTTP on teaching quality.

We also were interested to know whether teachers who had a math degree performed differently. To check whether accounting for those who majored in math would make a difference in our analysis, we examined the heterogeneous effect by including a new interaction term created by multiplying the treatment variable times the dummy variable (whether a teacher majored in math). The results are shown in Appendix Table A2, A3, and A4. Our analysis suggests that the NTTP has a positive heterogeneous effect on student academic performance, which is significant at the 5% level (Appendix Table A2, Column 2, Row 2). That is, if treatment teachers majored in math, the academic performance of their students is more likely to have been improved by the teachers' attendance at the NTTP. The heterogeneous effect of the NTTP on MKT test scores, however, is insignificant. Our interpretation of this result is that teachers who majored in math did not gain any teaching knowledge when they participated in the NTTP (Appendix Table A3, Column 2, Row 2). Ultimately, there is also no significant heterogeneous effect of the NTTP on any of the teaching practices of the treatment teachers who majored in math (Appendix Table A4, Column 2, Row 2).

There is one possible explanation for these surprising results. It may be that, in general, teachers who majored in math already have good training in how to teach math; thus, their students performed better in math than did the students of teachers who had not majored in math. In contrast, as they already had their own teaching knowledge system before the NTTP, when they participated in the NTTP, they may have received conflicting information that confused them, or they may have thought that they had enough teaching knowledge such that they did not make an effort to learn from the NTTP. Thus, teachers who majored in math did not increase their math teaching knowledge by participating in the NTTP, nor did they improve their teaching practices.

6. Conclusions

In general, we find that teacher participation in the NTTP had no positive impact on student achievement. We did not find a statistically significant difference in math achievement between treatment and control students in any of our models. In fact, teacher participation in the NTTP may even have lowered student academic achievement. Our findings also demonstrate that there were no significant heterogeneous effects.

We then sought to understand why the NTTP did not yield positive impacts. Based on our analysis, we can tentatively conclude that the lack of impact on student outcomes was not due to a lack of learning among the trained teachers. We find evidence to suggest that the teachers who received NTTP training had improved math knowledge for teaching relative to the control teachers.

This improvement in knowledge, however, did not result in any significant effect on teaching practices. If the NTTP is faulty in its design, it is likely because it fails to induce teachers to apply what they learned in the training within their classrooms. In addition, having their teachers attend the NTTP during the semester may have disrupted normal learning for the students. This mechanism may help to explain the negative impact of the NTTP on student academic achievement.

It should be noted that the study may suffer from a potential limitation because we examined changes in student achievement after only one semester. It could be that the change in teacher knowledge eventually will lead to greater gains for students over the course of a school year. In addition, it could be that future cohorts of students will benefit from the change in teacher knowledge. These hypothesized effects would nevertheless have to rely on teachers' having made a permanent change in the way that they teach. Given that teachers did not change their math teaching practices in the period of time immediately after participating in the study, the possibility of students' benefiting from a change in teacher knowledge may be limited.

The results of this study contribute to a broader policy debate about how to effectively invest in teacher PD programs, especially in rural China. Recently, there has been increasing support from officials in the Ministry of Finance of the People's Republic of China for greater investment in PD programs for rural teachers (MOE and MOF, 2015). Our results suggest that, if the government is interested in providing rural teachers with more opportunities to participate in the NTTP, such programs need to place a greater emphasis on helping teachers convert the skills and knowledge acquired during training into actual classroom practices.

There are programs in other countries that have such an emphasis. For example, 'clinical supervision,' which has been used as a method to foster teacher development through discussion, observation, and analysis of teaching 'in the clinic of the classroom' (Grimment and Crehan, 1992), has been used in several countries, including Pakistan (Garder 1995), Nigeria (Tatto 1997), Israel (Barak, Pearlman-Avnion, and Glanz 1997), and Brunei (Bourke 2001). More research is needed to determine the effectiveness of such programs. Similar classroom-based interventions, however, could offer the potential to improve the teaching quality and student educational outcomes in China.

If there are other PD programs that emphasize improving training effectiveness, the cost effectiveness of such programs is worthy of government consideration. Because the goal of the PD program is to improve the students' academic achievement, the measurement of effectiveness – the improvement of students' academic performance – is straightforward (Bartolic-Zlomislic and Bates 1999; Bartolic-Zlomislic and Brett 1999; Jung 2005) and can be examined by standardized tests. If the PD program has no or a negative impact on student academic achievement, as in the case of the NTTP, its effectiveness will be zero, and it can be considered cost ineffective. From a policy perspective, however, if the PD program has no or a negative impact on students' academic achievement, the program first should be made effective before there can be a discussion of cost effectiveness.

If the PD program can improve students' achievement, its cost effectiveness can be understood in terms of the cost of one standard-deviation gain in academic performance. Some studies have used a similar cost-per-unit standard to determine cost effectiveness (Jung 2005; Sylvia et al. 2015). A study conducted in Japan used the standard of a one-point gain in teacher learning achievement to show the cost effectiveness of teacher training programs (Jung 2005). Sylvia et al. (2015) used the standard of a line of visual acuity to estimate the cost effectiveness of a subsidized eyeglasses program.

Although the notion of cost effectiveness of the NTTP is meaningless, as it has no impact on student achievement, we would like to discuss the cost elements of the NTTP to assist in the further design of PD programs. We divide the costs of the NTTP into three categories – capital, operation, and opportunity – according to the commonplace cost categories of teacher training programs in the literature (Whalen and Wright 1999; Rumble 2001; Jung 2005). The capital costs of the NTTP are the costs for the training facility itself. Because the NTTP we evaluated is organized by prefectural primary schools, the programs used classrooms as training facilities instead of renting separate facilities, which means that were no capital costs for this program. There are, however, costs of using these facilities, such as electricity fees. Operation costs include trainer stipends, NTTP staff salaries, accommodations (as participating teachers were away from home and had to stay at a centralized location in the prefecture), and training materials. Opportunity costs are the time that participating teachers spent on the NTTP, which they could have spent preparing and teaching their classes, as well as teachers' travel costs, which they would not have if they did not participate in the NTTP. In regard to measuring the cost of the teacher trainings, the capital and operation costs are monetary costs that could be measured by looking at the expenditure of the NTTP office. The salaries of teachers could be used to measure the costs associated with the time that participating teachers spent on the NTTP. Travel costs could be calculated according to local transportation fees. It should be noted, however, that the cost elements we discuss above differ by PD program. PD programs with different designs will yield different costs (for example, classroom-based interventions program will incur the cost of a trainer who visits the classroom). Further studies are needed to determine the cost effectiveness of potential PD programs in China.

Notes

1. In addition to the NTTP, there are many other teacher PD programs that are run by local governments. As the nation's flagship program, the NTTP involves much larger expenditures per teacher and more prestige for participation than local teacher PD programs can offer.
2. There is no clear record of each schools' past participation in the NTTP; thus, we were unable to match control and treatment schools based on participation rates in the NTTP over time. Every school has the opportunity to gain slots in the NTTP each year in different subjects; for example, School A gains slots in the NTTP math program this year; School B gains slots in the NTTP Chinese program this year; and School C does not gain any slot this year but might already have gained slots last year or will gain slots next year. In other words, although the number of slots assigned to each school every year is different, on average, teachers of rural schools in the same county have a roughly equal opportunity to participate in the NTTP for the subject that they primarily teach. There is no specific policy, however, that clarifies this selection process, and we could not find consistent records of teacher participation by school. We received our information about school participation through interviews with the NTTP provincial officer. Based on this information and the three criteria noted above, we were able to effectively match control and treatment schools. We compared the difference of number of students and *tongkao* ranking between treatment schools and control schools (see Appendix Table A1). According to Appendix Table A1, there are no statistically significant differences between treatment schools and control schools in terms of these two characteristics. As to the distance between each pair of treatment and control schools, 64.71 percent of the control schools are in the same township as their paired treatment school. Although the remaining control schools are not in the same township as their paired treatment school, they are within a 30-minute drive from the paired treatment school. Therefore, our control schools are well matched with our treatment schools on these three criteria. Further, we checked for the balance on observable characteristics among the treatment and control groups and found that there was no imbalance significant enough to affect internal validity. Thus, we were able to successfully match control and treatment schools.
3. All treatment teachers who attended the math NTTP were math teachers. For this reason, we chose math teachers to fill all of the slots in the control group.
4. A left-behind child is defined as one whose parents left the home and migrated to urban areas for better job opportunities. These children are typically raised by close relatives, for example, paternal grandparents (Ye et al. 2006; Zhou et al. 2015).
5. In our study, a team of enumerators, who were trained and sent by our research team, conducted the student survey at both the baseline and endline. There were two enumerators for each class. When students were responding to the questionnaire, there were only students present (in addition to the two enumerators in the classroom). The two enumerators were responsible for not allowing others (including teachers) to come into the classroom when students were responding to the questionnaire. After students finished the questionnaire, the enumerators put all student questionnaires into a sealed bag. Hence, the sample teachers had no opportunity to either come into the classroom to examine the questionnaires while students were taking the survey or see the questionnaires after the students finished. Thus, the sample teachers had no knowledge of the second questionnaire block given to students or any

opportunity to influence the answers of the students. Therefore, it is highly unlikely that this could be a source of contamination.

6. In the GLS weighting procedure, a summarizing index is created from different dimensions. Variables (dimensions) with more variance in the control group are weighted more heavily, while variables with strong correlation are weighted less heavily. The index itself is, therefore, a weighted mean of these variables. For more information, see http://www.tandfonline.com/doi/pdf/10.1198/016214508000000841.
7. Our results show that the NTTP actually decreased the academic achievement of treatment students by 0.07 standard deviations, a result that was significant at the 10% level. From the literature, we believe that we can characterize a coefficient that measures a 0.07 standard-deviation difference as 'small' but not small enough to be considered random noise. In other published papers, authors have reported scores of that magnitude and interpret them to be 'significant.' For example, Rongeldt et al. (2013) shows that teacher turnover decreased student achievement by 0.06 to 0.08 standard deviations in a study that used the data of 85,000 New York City 4th- and 5th-grade students. Another study indicated that, among 2,430 students in an eastern province of China (Anhui), the English performance of students in private migrant schools was 0.1 standard deviations lower than that of students in public rural schools (Wang et al. 2017). The paper concluded that the quality of schooling in migrant schools was lower than that in rural public schools. Further, we can contextualize 0.07 standard deviations in the context of western rural China, the area of China from which we draw our sample. Specifically, the results of an experiment among 16,856 students in rural China suggests that computer-assisted learning in math improved student test scores by 0.1 standard deviations (which was significant at the 1% level; Mo et al. 2015). Another study showed that the effect of setting up a vision care program and providing glasses to students with myopia improved test scores by 0.25 standard deviations, with the point estimates equivalent to half a semester of additional learning (Ma et al. 2018). Compared with the findings on these papers, we believe that it is best to interpret 0.07 standard deviations as small but not equate it to noise.
8. In addition, we ran an analysis that considers the two types of control teachers (within-school and across-school). This analysis was conducted by adding a dummy variable that equals 1 if the teacher was a within-school control teacher, and 0 otherwise. The results of the NTTP on student academic achievement, using this approach, are substantively identical to those that we report in the manuscript. These results are reported in a supplemental appendix available online: (https://reap.fsi.stanford.edu/publication/impact-teacher-professional-development-programs-student-achievement-rural-china).
9. In fact, around a half standard-deviation gain in math knowledge is relatively small compared to gains reported in other studies of different short-term PD programs. One study of teacher PD programs in North Carolina, United States, found that a five-day training intervention resulted in a gain in math knowledge of 0.78 standard deviations among treatment teachers (Faulkner and Cain 2013). Another conducted in 10 sites across the United States found that an intensive two-week summer training program that combined facilitator guides and online support resources with in-person instruction increased math knowledge by 0.98 standard deviations among treatment teachers (Bell et al. 2010). It is important to note, however, that the first study examined an intervention that lasted four weeks longer than did the NTTP, while the second study investigated a relatively more intensive program (that would likely be much more expensive, perhaps even prohibitively so, to implement on as large a scale as China). To our knowledge, there are no studies that measure the effectiveness of PD in China to compare to our own results.
10. As part of our effort to better understand whether teachers apply their knowledge to their teaching practice, we conducted independent classroom observations immediately after the endline survey (February 2015). Our classroom observation approach was based on a modified protocol of the Classroom Assessment Scoring System (CLASS). The CLASS protocol provides a common framework to observe the quality of classrooms across grades (Pianta et al. 2007; Pianta, La Paro, and Hamre 2008). CLASS was designed to observe and document instructional support, emotional support, classroom organization, and student engagement in the classroom (Pianta et al. 2007; Pianta, La Paro, and Hamre 2008). In our carrying out of CLASS-based observations in our study schools, a team of two enumerators conducted the observations in eight randomly selected classrooms (four treatment classrooms and four control classrooms). Before conducting the observations, the enumerators participated in a step-by-step instructional training course on how to conduct classroom observations. When enumerators conducted the observations, they sat in the back of the classroom, observed, and recorded the classroom practice in terms of certain behavioral, emotional, and physical markers without any intervention into or disturbance of the class. During the observation period, the two enumerators independently recorded their individual observations without communicating with each other. After each class, they compared their records and decided on the final observation results.

Acknowledgments

The authors would like to acknowledge the funding of the 111 project (高等学校学科创新引智计划资助, B16031).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the 111 project (高等学校学科创新引智计划资助) [B16031] and Gratudate Student Innovation Fund of Shannxi Normal Univeristy [2017CSY056].

Notes on contributors

Meichen Lu is a Ph.D student at the Department of Agricultural and Applied Economics, University of Georgia. Her research has been focused on development economics, rural management and economics of education. This work was undertaken when she was working as a research assistant at Center for Experimental Economics in Education of Shaanxi Normal University.

Prashant Loyalka is an Assistant Professor at the Graduate School of Education of Stanford University and a Center Fellow at the Freeman Spogli Institute for International Studies. His research focuses on addressing inequalities in the education of youth and on improving the quality of education received by youth in multiple countries including China, India, Russia, and the United States.

Yaojiang Shi is a Professor and the founding director of the center for Experimental Economics in Education, Shaanxi Normal University in China. His research focuses on the emergence and development of economic enterprises in rural China and China's education reforms. He uses empirical research to identify important leverage points for education policy that addressed the needs of the rural poor.

Fang Chang is an assistant Professor at the center for Experimental Economics in Education, Shaanxi Normal University in China. Her research has been in the field of rural development as well as rural education, with focus on teacher professional development and teacher performance payment.

Chengfang Liu is an associate Professor with tenure at the China Center for Agricultural Policy, School of Advanced Agricultural Science, Peking University. Her research has been in the field of agricultural and rural development, with focus on the provision of rural infrastructure, human capital, and migration.

Scott Rozelle is the Helen F. Farnsworth Senior Fellow and the co-director of the Rural Education Action Program in the Freeman Spogli Institute for International Studies at Stanford University. His research focuses on China, especially agricultural policy, the emergence and evolution of markets and other economic institutions, and the economics of poverty and inequality.

References

Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (1): 95–135. doi:10.1086/508733.

Agodini, R., B. Harris, S. Atkins-Burnett, S. Heaviside, T. Novak, and R. Murphy. 2009. *Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools. NCEE 2009-4052*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Anderson, M. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Revaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103: 1481–1495. doi:10.1198/016214508000000841.

Ball, D. L., and D. K. Cohen. 1999. "Developing Practice, Developing Practitioners: Toward a Practice-Based Theory of Professional Education. Teaching as the Learning Profession." *Handbook of Policy and Practice* 1: 3–22.

Barak, M., S. Pearlman-Avnion, and J. Glanz. 1997. "Using Developmental Supervision to Improve Science and Technology Instruction in Israel." *Journal of Curriculum and Supervision* 12 (4): 367.

Bartolic-Zlomislic, S., and A. W. Bates, 1999. "Assessing the Costs and Benefits of Telelearning: A Case Study from the University of British Columbia." Network of Centers of Excellence (NCE)-Telelearning Project report.

Bartolic-Zlomislic, S., and C. L. A. R. E. Brett, 1999. Assessing the costs and benefits of telelearning: A case study from the Ontario Institute for Studies in Education of the University of Toronto. Retrieved February, 25, p.2005.

Bell, A., M. Wilson, T. Higgins, and B. McCoach. 2010. "Measuring the Effects of Professional Development on Teacher Knowledge: The Case of Developing Mathematical Ideas." *Journal for Research in Mathematics Education* 41 5: 479–512.

Borko, H. 2004. “Professional Development and Teacher Learning: Mapping the Terrain.” *Educational Researcher* 33 (8): 3–15. doi:10.3102/0013189X033008003.

Bourke, M. 2001. “The Role of the TP [Teaching Practice] TESL [Teaching English as a Second Language] Supervisor.” *Journal of Education for Teaching: International Research and Pedagogy* 27 (1): 63–73. doi:10.1080/02607470120042546.

Bruns, B., and J. Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: World Bank.

Chapman, W., X. Chen, and A. Postiglione. 2000. “Is Preservice Teacher Training Worth the Money? A Study of Teachers in Ethnic Minority Regions of the People’s Republic of China.” *Comparative Education Review* 44 (3): 300–328. doi:10.1086/447616.

China National Bureau of Statistics (CNBS). 2014. *China Statistical Yearbook*. Beijing: China Statistics Press.

Chu, H., P. Loyalka, J. Chu, Q. Qu, Y. Shi, G. Li, and S. Rozelle. 2015. “The Impact of Teacher Credentials on Student Achievement in China.” *China Economic Review* 36: 14–24. doi:10.1016/j.chieco.2015.08.006.

Cobb, J. 2000. “The Impact of a Professional Development School on Preservice Teacher Preparation, Inservice Teachers’ Professionalism, and Children’s Achievement: Perceptions of Inservice Teachers.” *Action in Teacher Education* 22 (3): 64–76. doi:10.1080/01626620.2000.10463021.

Cobb, L. 1999. *An International Comparison of Teacher Education. ERIC Digest*. Washington DC: ERIC Clearinghouse on Teaching and Teacher Education.

Cohen, K., and C. Hill. 2000. “Instructional Policy and Classroom Performance: The Mathematics Reform in California.” *Teachers College Record* 102 (2): 294–343. doi:10.1111/tcre.2000.102.issue-2.

Copur-Gencturk, Y., and T. Lubienski. 2013. “Measuring Mathematical Knowledge for Teaching: A Longitudinal Study Using Two Measures.” *Journal of Mathematics Teacher Education* 16 (3): 211–236. doi:10.1007/s10857-012-9233-0.

Dadds, M. 2001. “Continuing Professional Development: Nurturing the Expert Within.” In *Teacher Development: Exploring Our Own Practice*, edited by J. Soler, A. Craft, and H. Burgess. London: Paul Chapman Publishing and The Open University: 50-57.

Darling-Hammond, L. 2000. “Teacher Quality and Student Achievement: A Review of State Policy 2000.” *Education Policy Analysis* 8: 1. doi:10.14507/epaa.v8n1.2000.

Darling-Hammond, L., and W. McLaughlin. 1995. “Policies that Support Professional Development in an Era of Reform.” *Phi Delta Kappan* 76 (8): 597.

Delaney, S., L. Ball, C. Hill, G. Schilling, and D. Zopf. 2008. “Mathematical Knowledge for Teaching: Adapting US Measures for Use in Ireland.” *Journal of Mathematics Teacher Education* 11 (3): 171–197. doi:10.1007/s10857-008-9072-1.

Evans, K., and A. Popova. 2016. “What Really Works to Improve Learning in Developing Countries? an Analysis of Divergent Findings in Systematic Reviews.” *The World Bank Research Observer* 31 (2): 242–270. doi:10.1093/wbro/lkw004.

Faulkner, N., and R. Cain. 2013. “Improving the Mathematical Content Knowledge of General and Special Educators Evaluating a Professional Development Module that Focuses on Number Sense. Teacher Education and Special Education.” *The Journal of the Teacher Education Division of the Council for Exceptional Children* 36 (2): 115–131.

Fishman, J., W. Marx, S. Best, and T. Tal. 2003. “Linking Teacher and Student Learning to Improve Professional Development in Systemic Reform.” *Teaching and Teacher Education* 19 (6): 643–658. doi:10.1016/S0742-051X(03)00059-3.

Fryer, R. G. 2017. “The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments.” *Handbook of Economic Field Experiments* 2: 95–322.

Garder, R. 1995. “On-Service Teacher Education.” In *International Encyclopedia of Teaching and Teacher Education*, edited by L. Anderson. Second ed. London: Pergamon Press: 628-632.

Garet, M.S., S. Cronen, M. Eaton, A. Kurki, M. Ludwig, W. Jones, K. Uekawa, A. Falk, H. S. Bloom, F. Doolittle, and P. Zhu. 2008. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*. Washington, DC: U.S. Department of Education, NCEE (2008). 2008–2403.

Garet, S., C. Porter, L. Desimone, F. Birman, and S. Yoon. 2001. “What Makes Professional Development Effective? Results from a National Sample of Teachers.” *American Educational Research Journal* 38 (4): 915–945. doi:10.3102/00028312038004915.

Government of Chile. 2003. “Attracting, Developing, and Retaining Effective Teachers: OECD Activity, Country Background Report for Chile.” Santiago: http://www.oecd.org/chile/26742861.pdf

Government of India. 2013. “Innovative Teacher Training Modules.” New Delhi: http://www.teindia.nic.in/Te_Trg_Module.aspx

Grimmett, P., and P. Crehan. 1992. “The Nature of Collegiality in Teacher Development: The Case of Clinical Supervision.” *Teacher Development and Educational Change* 12 (2): 56–85.

Heller, J. I., K. R. Daehler, N. Wong, M. Shinohara, and L. W. Miratrix. 2012. “Differential Effects of Three Professional Development Models on Teacher Knowledge and Student Achievement in Elementary Science.” *Journal of Research in Science Teaching* 49 (3): 333–362. doi:10.1002/tea.21004.

Hiebert, J., and D. A. Grouws. 2007. "The Effects of Classroom Mathematics Teaching on Students' Learning." *Second Handbook of Research on Mathematics Teaching and Learning* 1: 371–404.
Hill, C., B. Rowan, and L. Ball. 2005. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Journal* 42 (2): 371–406. doi:10.3102/00028312042002371.
Hill, C., G. Schilling, and L. Ball. 2004. "Developing Measures of Teachers' Mathematics Knowledge for Teaching." *Elementary School Journal* 105: 11–30. doi:10.1086/428763.
Huang, A., and X. Du. 2007. "Comparative Analysis of Urban and Rural Difference in Family Education in China." *Journal of Yibing University* 1: 107–110. (In Mandarin Chinese).
Jacob and Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." *Journal of Human Resources* 39 (1): 50–79. 2004. doi:10.2307/3559005.
Jung, I. 2005. "ICT-pedagogy Integration in Teacher Training: Application Cases Worldwide." *Educational Technology & Society* 8 (2): 94–101.
Kane, T. J., and D. O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." (*No. W14607*).Cambrige, MA: National Bureau of Economic Research.
King, E. M., and J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *The World Bank Research Observer* 24 (1): 55–82. doi:10.1093/wbro/lkn009.
Kuziemko, I. 2006. "Using Shocks to School Enrollment to Estimate the Effect of School Size on Student Achievement." *Economics of Education Review* 25 (1): 63–75. doi:10.1016/j.econedurev.2004.10.003.
Li, F., P. Loyalka, C. Liu, H. Yi, L. Zhang, J. Chu, N. Johnson, and S. Rozelle, 2015, "The Impact of Conditional Cash Transfers on the Matriculation of Junior High School Students into Rural China's High Schools." Working Paper, The Rural Education Action Program, Freeman Spogli Institute for Foreign Policy, Stanford, CA. Available at SSRN 2610584
Li, H., P. Loyalka, S. Rozelle, and B. Wu. 2017. "Human Capital and China's Future Growth." *Journal of Economic Perspectives* 31 (1): 25–48.
Liu, C., L. Zhang, R. Luo, S. Rozelle, B. Sharbono, and Y. Shi. 2009. "Development Challenges, Tuition Barriers and High School Education in China." *Asia Pacific Journal of Education* 29 (4): 503–520. doi:10.1080/02188790903312698.
Liu, H., C. C. F. Liu, and P. Loyalka. 2016. "Describing the Landscape of Teacher Training in China." *China & World Economy* 24 (3): 86–104. doi:10.1111/cwe.12160.
Loyalka, P., C. Liu, Y. Song, H. Yi, X. Huang, J. Wei, and S. Rozelle. 2013. "Can Information and Counseling Help Students from Poor Rural Areas Go to High School? Evidence from China." *Journal of Comparative Economics* 41 (4): 1012–1025. doi:10.1016/j.jce.2013.06.004.
Loyalka, P., J. Chu, J. Wei, N. Johnson, J. Reniker, and S. Rozelle, 2014. "Inequalities in the Pathway to College in China: When Do Students from Poor Areas Fall Behind?" Working Paper, The Rural Education Action Program, Freeman Spogli Institute for Foreign Policy, Stanford, CA.
Loyalka, P., S. Sylvia, C. Liu, J. Chu, and Y. Shi 2016. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." Working Paper, The Rural Education Action Program, Freeman Spogli Institute for Foreign Policy, Stanford, CA.
Luo, R., Y. Shi, L. Zhang, C. Liu, S. Rozelle, and B. Sharbono. 2009. "Malnutrition in China's Rural Boarding Schools: The Case of Primary Schools in Shaanxi Province." *Asia Pacific Journal of Education* 29 (4): 481–501. doi:10.1080/02188790903312680.
Ma, Y., N. Congdon, Y. Shi, R. Hogg, A. Medina, M. Boswell, S. Rozelle, and M. Iyer. 2018. "Effect of A Local Vision Care Center on Eyeglasses Use and School Performance in Rural China: A Cluster Randomized Clinical Trial." *JAMA Ophthalmology* 136 (7): 731–737. doi:10.1001/jamaophthalmol.2018.1329.
Metzler, J., and L. Woessmann. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." *Journal of Development Economics* 99 (2): 486–496. doi:10.1016/j.jdeveco.2012.06.002.
Ministry of Education (MOE). 2011. "Chinese National Curriculum Framework." Government of China, Accessed 2 September 2014: http://www.moe.gov.cn/publicfiles/business/htmlfiles/moe/s8001/201404/xxgk_167340.html
Ministry of Education (MOE). 2012. "Training Framework of National Teacher Training Program." Government of China: http://old.moe.gov.cn/ewebeditor/uploadfile/2012/10/12/20121012170215824.pdf doi:10.1094/PDIS-11-11-0999-PDN.
Ministry of Education and Ministry of Finance (MOE and MOF). 2010. "Notice of the Implementation of the National Training Plan for Primary and Secondary Education Teachers." Government of China http://www.gov.cn/zwgk/2010-06/30/content_1642031.htm
Ministry of Education and Ministry of Finance (MOE and MOF). 2015. "Notice of Reforming the National Training Plan for Pre-School, Primary and Secondary Education Teachers." Government of China http://www.mof.gov.cn/zhengwuxinxi/zhengcefabu/201509/t20150907_1452338.htm
Mo, D., L. Zhang, H. Yi, R. Luo, and S. Brinton. 2013. "School Dropouts and Conditional Cash Transfers: Evidence from a Randomized Controlled Trial in Rural China's Junior High Schools." *Journal of Development Studies* 49 (2): 190–207. doi:10.1080/00220388.2012.724166.

Mo, D., W. Huang, Y. Shi, L. Zhang, M. Boswell, and S. Rozelle. 2015. "Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China." *China Economic Review* 36: 131–145. doi:10.1016/j.chieco.2015.09.001.
Muralidharan, K. 2012. *Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India*. Society for Research on Educational Effectiveness, India.
Muralidharan, K., and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *The Journal of Political Economy* 119 (1): 39–77. doi:10.1086/659655.
OECD. 2014. *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science*. Paris, France: PISA, OECD Publishing. doi:http://dx.doi.org/10.1787/9789264201118-en.
Olakulehin, K. 2007. "Information and Communication Technologies in Teacher Training and Professional Development in Nigeria." *Turkish Online Journal of Distance Education* 8 (1): 133–142.
Overbaugh, R., and R. Lu. 2008. "The Impact of a NCLB-EETT Funded Professional Development Program on Teacher Self-Efficacy and Resultant Implementation." *Journal of Research on Technology in Education* 41 (1): 43–61. doi:10.1080/15391523.2008.10782522.
Owston, R., H. Wideman, J. Murphy, and D. Lupshenyuk. 2008. "Blended Teacher Professional Development: A Synthesis of Three Program Evaluations." *The Internet and Higher Education* 11 (3): 201–210.
Pianta, C., K. Hamre, J. Haynes, L. Mintz, and M. La Paro. 2007. *Classroom Assessment Scoring System Manual, Middle/Secondary Version*. Charlottesville: University of Virginia.
Pianta, C., M. La Paro, and K. Hamre. 2008. *Classroom Assessment Scoring System, Manual, K-3*. Baltimore, MD: Brookes.
Rivkin, S., E. Hanushek, and J. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–458. doi:10.1111/ecta.2005.73.issue-2.
Rockoff, J. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *The American Economic Review* 94 (2): 247–252.
Ronfeldt, M., S. Loeb, and J. Wyckoff. 2013. "How Teacher Turnover Harms Student Achievement." *American Educational Research Journal* 50 (1): 4–36. doi:10.3102/0002831212463813.
Rumble, G. 2001. "The Costs and Costing of Networked Learning." *Journal of Asynchronous Learning Networks* 5 (2): 75–96.
Saxe, G. B., and M. Gearhart. 2001. "Enhancing Students' Understanding of Mathematics: A Study of Three Contrasting Approaches to Professional Support." *Journal of Mathematics Teacher Education* 4 (1): 55–79. doi:10.1023/A:1009935100676.
Schifter, D., S. J. Russell, and V. Bastable. 1999. "Teaching to Big Ideas." In *The Diagnostic Teacher: Constructing New Approaches to Professional Development*, edited by M. Z. Solomon. New York: Teacher College Press: 22-47.
Shaanxi statistical bureau. 2014. *Shaanxi Statistical Yearbook*. Beijing: China Statistics Press.
Shi, Y., L. Zhang, Y. Ma, H. Yi, C. Liu, N. Johnson, and S. Rozelle. 2015. "Dropping Out of Rural China's Secondary Schools: A Mixed-Methods Analysis." *The China Quarterly* 224: 1048–1069. doi:10.1017/S0305741015001277.
Showers, B., B. Joyce, and B. Bennett. 1987. "Synthesis of Research on Staff Development: A Framework for Future Study and A State-Of The-Art Analysis." *Educational Leadership* 45 (3): 77–87.
Sylvia, S., X. Ma, Y. Shi, S. Rozelle, and C. Lawell, 2015. "Ordeal Mechanisms, Information and the Cost-Effectiveness of Subsidies: Evidence from Subsidized Eyeglasses in Rural China." Working Paper, The Rural Education Action Program, Freeman Spogli Institute for Foreign Policy, Stanford, CA.
Tatto, T. 1997. "Limits and Constraints to Effective Teacher Education." In *International Handbook of Education and Development: Preparing Schools, Students and Nations for the Twenty-First Century*, edited by Cummings, W.K and McGinn, N.F. Lodon: Pergamon Press: 213–230.
UNESCO Institute for Statistics. 2006. *Teachers and Educational Quality: Monitoring Global Needs for 2015*. Montreal: International Institute for Educational Planning, UNES.
Vegas, E. 2007. "Teacher Labor Markets in Developing Countries." *The Future of Children* 17 (1): 219–232.
Villegas-Reimers, E. 2003. *Teacher Professional Development: An International Review of the Literature*. Paris: International Institute for Educational Planning.
Wang, H., C. Yang, F. He, Y. Shi, Q. Qu, S. Rozelle, and J. Chu. 2015. "Mental Health and Dropout Behavior: A Cross-Sectional Study of Junior High Students in Northwest Rural China." *International Journal of Educational Development* 41: 1–12. doi:10.1016/j.ijedudev.2014.12.005.
Wang, X., C. Liu, L. Zhang, R. Luo, Y. Shi, S. Rozelle, and B. Sharbono. 2009. "What Is Keeping the Poor Out of College? Enrollment Rates, Educational Barriers and College Matriculation in China." *China Agricultural Economic Review* 3: 131–149. doi:10.1108/17561371111131281.
Wang, X., R. Luo, L. Zhang, and S. Rozelle. 2017. "The Education Gap of China's Migrant Children and Rural Counterparts." *The Journal of Development Studies* 53 (11): 1865–1881. doi:10.1080/00220388.2016.1274395.
Whalen, T., and D. Wright. 1999. "Methodology for Cost-Enefit Analysis of Web-Based Tele-Learning: Case Study of the Bell Online Institute." *American Journal of Distance Education* 13 (1): 24–44. doi:10.1080/08923649909527012.
Wilson, S. M., and J. Berne. 1999. "Chapter 6: Teacher Learning and the Acquisition of Professional Knowledge: An Examination of Research on Contemporary Professional Development." *Review of Research in Education* 24 (1): 173–209.

World Bank. 2001. *China: Challenges of Secondary Education*. Second Education Series 22856. Washington D.C: World Bank.

Yan, H., F. Wei, and B. Li. 2013. "The Situation and Development of a Roadmap of Teacher Training." *Teacher Education* 5 (6): 43–50.

Ye, J., and L. Pang. 2011. "Differentiated Childhoods: Impacts of Rural Labor Migration on Left-Behind Children in China." *The Journal of Peasant Studies* 38 (2): 355–377.

Ye, J., Y. Wang, K. Zhang, and J. Lu. 2006. "Life Affection of Parents Going Out for Work to Left behind Children." *Chinese Rural Econ* 1: 57–65. (in Chinese).

Yi, H., L. Zhang, R. Luo, Y. Shi, D. Mo, X. Chen, C. Brinton, and S. Rozelle. 2012. "Dropping Out: Why are Students Leaving Junior High in China's Poor Rural Areas?." *International Journal of Educational Development* 32 (4): 555–563. DOI:10.1016/j.ijedudev.2011.09.002.

Yoon, S., T. Duncan, Y. Lee, B. Scarloss, and L. Shapley, 2007. "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement." Issues & Answers.REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1). doi:10.1094/PDIS-91-4-0467B.

Yoshikawa, H., D. Leyva, C. E. Snow, E. Treviño, M. Barata, C. Weiland, and M. C. Arbour. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Developmental Psychology* 51 (3): 309. doi:10.1037/a0038785.

Zhang, L., F. Lai, X. Pang, H. Yi, and S. Rozelle. 2013. "The Impact of Teacher Training on Teacher and Student Outcomes: Evidence from a Randomized Experiment in Beijing Migrant Schools." *Journal of Development Effectiveness* 5 (3): 339–358. doi:10.1080/19439342.2013.807862.

Zhou, C., S. Sylvia, L. Zhang, R. Luo, H. Yi, C. Liu, Y. Shi, et al. 2015. "China's Left-Behind Children: Impact of Parental Migration on Health, Nutrition, and Educational Outcomes." *Health Affairs* 34 (11): 1964–1971. doi:10.1377/hlthaff.2015.0150.

Zuo, E., and Q. Su. 2012. "The Investigation of Rural Middle School Core Physical Education Teachers Participating in Hunan Province 'National Training Program'." *Nenjiang Technology* 3(2): 20–25. (in Chinese).

Table A1. Comparisons between treatment schools and control schools between treatment schools and control schools.

Variable	Treatment school (1) Mean Sd	Control school (2) Mean Sd	Difference: (1) – (2) (3)	P-value (4)
1. Number of students	593.35 [0.64]	581.88 [0.68]	11.47 (45.24)	0.80
2. Tongkao ranking (1 = above median; 0 = equal or below median)	0.57 [0.49]	0.60 [0.50]	−0.30 (0.06)	0.72
Number of observations	34	34	68	

Source: Authors' survey. *** p < 0.01, ** p < 0.05, * p < 0.1.
Note: Tongkao ranking is conducted at the county level, and students in each county were given the same standard tests. Thus, schools with similar tongkao ranking reflect similar teaching and teacher quality to some extent

Table A2. Heterogeneous effects of NTTP on student academic achievement: by whether or not teacher majored in math.

Dependent variable: Endline student math score (SD)	(1)	(2)
1.Teacher participates in teacher training (1 = yes, 0 = no)	−0.05	−0.11**
	(0.05)	(0.05)
2. Teacher participates in teacher training * Teacher majored in math	0.14	0.24**
	(0.10)	(0.11)
3. Teacher majored in math (1 = yes, 0 = no)	−0.08	−0.09
	(0.06)	(0.07)
4. Baseline student math score (SD)	0.53***	0.52***
	(0.02)	(0.02)
Student characteristics	NO	YES
Teacher characteristics	NO	YES
School-pair fixed effect	YES	YES
Constant	0.05	1.67***
	(0.04)	(0.28)
Observations	3066	3066
R-squared	0.398	0.414

Source: Authors' survey. *** p < 0.01, ** p < 0.05, * p < 0.1.
Notes:
(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.
(b) The student characteristics include: whether the student lives at school, whether the student's mother completed junior high school, and whether the student's father completed junior high school.
(c) The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, teaching experience whether teacher undergone provincial PD in the last year and whether teacher undergone prefecture PD in the last year.
(d) The heterogeneous effects analyses include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.

Table A3. Heterogeneous effects of NTTP on math knowledge for teaching (MKT) of teachers: by whether or not teacher majored in math.

Dependent variable: Endline teaching practice index (SD) (Report from Students)	(1)	(2)
1.Teacher participates in teacher training (1 = yes, 0 = no)	0.78***	0.61**
	(0.21)	(0.23)
2. Teacher participates in teacher training * Teacher majored in math	−1.73***	−1.05
	(0.45)	(0.63)
3. Teacher majored in math (1 = yes, 0 = no)	0.23	−0.11
	(0.35)	(0.45)
4. Baseline teacher MKT test score (SD)	0.61***	0.61***
	(0.13)	(0.14)
Teacher characteristics	NO	YES
School-pair fixed effect	NO	YES
Constant	YES	YES
	−0.27*	−0.23
Observations	(0.15)	(0.48)
R-squared	82	82

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes:

(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.

(b) The unadjusted model does not control for teacher characteristics while the adjusted model does. The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, teaching experience whether teacher undergone provincial PD in the last year and whether teacher undergone prefecture PD in the last year.

(c) The unadjusted model and adjusted model both include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.

(d) Two teachers (one that participated in the NTTP and one that did not participate in the NTTP) were unable to complete the teacher MKT test in the endline survey. Thus we have 82 (instead of the 84 in our sample) teachers in this analysis.

Table A4. Heterogeneous effects of NTTP on the teaching practices of teachers: by whether or not teacher majored in math.

Dependent variable: Endline teaching practice index (SD) (Report from Students)	(1)	(2)
1.Teacher participates in teacher training (1 = yes, 0 = no)	0.02	0.01
	(0.04)	(0.04)
2. Teacher participates in teacher training * Teacher majored in math	−0.02	−0.08
	(0.08)	(0.11)
3. Teacher majored in math (1 = yes, 0 = no)	−0.02	0.03
	(0.06)	(0.07)
4. Baseline teacher teaching practice index (SD)	0.32***	0.32***
	(0.03)	(0.02)
Teacher characteristics	NO	YES
School-pair fixed effect	YES	YES
Constant	0.00	−0.05
	(0.03)	(0.06)
Observations	3,066	3,066
R-squared	0.164	0.173

Source: Authors' survey. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes:

(a) Cluster-robust standard errors adjusted for clustering at the student-teacher level in parentheses.

(b) The teacher characteristics include: teacher gender, whether the teacher completed university, whether the teacher majored in math, whether the teacher has achieved the highest teaching rank, teaching experience whether teacher undergone provincial PD in the last year and whether teacher undergone prefecture PD in the last year.

(c) The heterogeneous effects analyses include a school-pair fixed effect. Since there were 34 pairs of treatment schools and control schools in our sample, we include a matrix made up of 34 school pair dummy variables.