December 2018

Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program

Prashant Loyalka, Anna Popova, Guirong Li, and Zhaolei Shi

Abstract

Despite massive investments in teacher professional development (PD) programs in developing countries, there is little evidence on their effectiveness. We present results of a large-scale, randomized evaluation of a national PD program in China in which teachers were randomized to receive PD; PD plus follow-up; PD plus evaluation of the command of PD content; or no PD. Precise estimates indicate PD and associated interventions failed to improve teacher and student outcomes after one year. A detailed analysis of the causal chain shows teachers find PD content to be overly theoretical, and PD delivery too rote and passive, to be useful.

Working Paper 330

August 2018

reap.fsi.stanford.edu

REAP Rural Education Action Program

Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program

By PRASHANT LOYALKA, ANNA POPOVA, GUIRONG LI, AND ZHAOLEI SHI*

Despite massive investments in teacher professional development (PD) programs in developing countries, there is little evidence on their effectiveness. We present results of a large-scale, randomized evaluation of a national PD program in China in which teachers were randomized to receive PD; PD plus follow-up; PD plus evaluation of the command of PD content; or no PD. Precise estimates indicate PD and associated interventions failed to improve teacher and student outcomes after one year. A detailed analysis of the causal chain shows teachers find PD content to be overly theoretical, and PD delivery too rote and passive, to be useful.

* Loyalka: Graduate School of Education, Stanford University, E413 Encina Hall, Stanford, CA 94305 (loyalka@stanford.edu), Popova: Graduate School of Education, Stanford University, 69 Cubberley 485 Lasuen Mall, Stanford, CA 94305 (apopova@stanford.edu), Li: School of Education, Henan University, Kaifeng, Henan, 475001 (guirong1965@163.com), Shi: Stanford University, 485 Lasuen Mall, Stanford, CA 94305 (henryzhaoleishi@gmail.com). Corresponding Author: Li. We would like to thank Alex Eble, David Evans, Rob Fairlie, Erik Hanushek, Scott Rozelle, and Sean Sylvia for helpful comments on an earlier version of the manuscript. This research was approved by the Institutional Review Board (IRB) through Stanford University (protocol number 35602).

Teacher quality is important for improving student achievement (Rockoff 2004; Hanushek and Rivkin 2010; Chetty, Friedman, and Rockoff 2014; Bruns and Luque 2015). For example, the difference between a high and low quality teacher amounts to a difference of 0.36 standard deviations (SDs) in student test scores in Uganda (Buhl-Wiggers et al. 2017), and a 0.54 SD difference in Pakistan (Bau and Das 2017). Unfortunately, researchers have found that a large proportion of teachers in developing countries are ill-prepared for teaching, lacking the requisite knowledge and skills to improve student achievement (Behrman et al. 1997; Villegas-Reimers 1998; Ball 2000; Behrman, Ross, and Sabot 2008; Bruns and Luque 2015; Tandon and Fukao 2015; Bold et al. 2017). Despite sometimes high levels of formal education among teachers, many exhibit weak cognitive skills and ineffective classroom practice (Bruns and Luque 2015; Bold et al. 2017).

Aware of the role that teacher quality can play in improving student learning outcomes, policymakers from developing countries have, like their counterparts in developed countries, established teacher professional development (PD) programs (Cobb 1999; Villegas-Reimers 2003; Vegas 2007). The aims of teacher PD programs are to help teachers gain subject-specific knowledge and skills, use appropriate instructional practices, and develop positive attitudes and values (all of which have strong positive associations with student achievement—Schifter, Russell and Bastable 1999; Villegas-Reimers 2003; Hill, Rowan and Ball 2005; Hiebert and Grouws 2007; Darling-Hammond and McLaughlin 2011; Metzler and Woessman 2012; Shepherd 2015; Bold et al. 2017). Despite these positive aims, teacher PD programs may fail to improve teacher and student outcomes if their content is of poor quality or limited relevance, if they are delivered ineffectively, if they lack follow-up to help teachers translate learning into practice, or if the programs fail to hold trainees accountable for their teaching (Subirats and Nogales 1989; Cohen 1990; Silva 1991; Braslavsky and Birgin 1992; Lieberman 1995; Corcoran 1995; Guskey 1995; Villegas-Reimers 1998; Schifter, Russell and

Bastable 1999; Ganser 2000; Villegas-Reimers 2003). Moreover, since teacher PD programs further require teachers, school administrators and policymakers to substitute teacher time and resources away from students, they may even lead to negative impacts.

Despite their purported importance, evidence on the effectiveness of PD programs has hitherto been lacking (OECD 2009; Bruns and Luque 2015). There have been almost no large-scale randomized evaluations of teacher PD programs on student achievement in developing countries.[1] High quality evidence from developed countries is also sparse. For example, a recent review of PD in mathematics identified over 600 studies of math PD interventions, of which only five were high-quality randomized control trials (Gersten et al. 2014). Another recent U.S.-focused review of PD more broadly, identified 1,300 PD studies, of which only nine had pre- and post-test data and a control group (Yoon et al. 2007).[2] While recent reviews appear to show that teacher PD requires significant contact hours, a detailed implementation plan, and ample follow-up support to be effective (Yoon et al. 2007; Fryer 2016), there is considerable variation in the size and direction of effects across reviewed studies. The absence of rigorous evidence on what works hampers the ability of policymakers to effectively invest in teacher PD programs (as well as determine how much to invest) and improve the quality of education systems.

[1] The only exception in a developing country context is Yoshikawa et al. (2015) who use a cluster randomized design to assess the impacts of a pilot PD program for early childhood education teachers in 64 schools in Chile. Yoshikawa et al. find moderate impacts on emotional and instructional support and classroom organization, but no impacts on student outcomes. Our evaluation further contributes to the literature by being the first to focus on teacher training in K-12 education specifically, as well as the first to evaluate an existing national policy similar to those implemented in many countries, as opposed to a researcher-designed pilot program.

[2] These experimental studies drew on small samples of only 5 to 44 teachers, and the PD programs they evaluated were implemented by the individuals who developed them, limiting their policy-relevance (Garet et al. 2011). Even the most rigorous developed country evaluations seem to have limited statistical power. For example, two recent experimental evaluations by the U.S. Department of Education of PD for early reading instruction and middle school both found no significant impacts on student achievement, but these findings drew on only 30 schools in each treatment group, and a sample of just 92 teachers, respectively (Garet et al. 2008; Garet et al. 2011).

Given these gaps in knowledge, the overall purpose of this paper is to evaluate the impact of a typical, large-scale teacher PD program from a developing country on a wide range of teacher and student outcomes. As secondary objectives, we endeavor to understand which types of students and teachers are impacted by teacher PD programs and why teacher PD programs may or may not be effective. Since one of the major purposes of teacher PD programs is to create a core group of teachers that can influence the teaching practices of other teachers (Darling-Hammond, Bullmaster, and Cobb 1995; Cochran-Smith and Lytle 1999; Berry 2011; Zepeda 2012), we also examine the degree to which PD programs have positive spillovers on peer teachers and students.[3]

To fulfill these goals, we conducted a large randomized evaluation of the Chinese government's flagship National Teacher Training Program (NTTP), which shares many characteristics typical of at-scale PD programs across low- and middle-income countries (Popova et al. 2018). Specifically, the NTTP has a focus on subject-specific content and pedagogy, is delivered primarily through lectures and discussion, and involves a block of initial face-to-face training with few follow-up visits. We also evaluate the effects of two accompanying post-training interventions that are believed to strengthen the impact of teacher PD. The post-training interventions consist of: (a) continuous *follow-up* with trainees (about supplementary materials, assignments, and progress reports); and (b) an *evaluation* of how much trainees recalled from the PD program. In conducting the evaluations, we analyze survey data on 600 teachers and 33,492 students in 300 schools, as well as extensive observational and interview data from a large number of teachers, their PD sessions, and their classrooms.

[3] Positive spillovers may be likely in countries such as China where teachers have frequent opportunities to interact and observe each other teaching in professional learning communities at schools (Sargent 2015). This is especially true in rural schools where the number of teachers is small.

We present five main sets of results. First, we find that neither teacher PD alone nor teacher PD with follow-up and/or evaluation have significant impacts on achievement after one academic year. Our study is sufficiently powered to identify even small effects (0.11 SDs and greater), meaning that the null findings should be taken seriously. Second, we find virtually no impacts on a wide range of secondary outcomes that would suggest impacts on student achievement could arise in the longer term. For example, no combination of PD with or without post-training follow-up or evaluation has significant impacts on subject-specific psychological factors among students, such as math anxiety or motivation, nor on teacher knowledge, attitudes, or teaching practices. As such, it is unlikely that the lack of impact on student achievement is due to the length of our evaluation timeframe. Third, and unsurprisingly given the absence of direct effects, we find no spillover effects of PD on students whose teachers did not receive PD. Fourth, using qualitative and quantitative data to further explore mechanisms, we propose two major reasons that may explain the lack of impacts: (a) the content of PD is overly theoretical and hard for teachers to implement; (b) the delivery of PD content is rote and passive, making it difficult for teachers to remember and relate to.

Finally, we consider heterogeneous effects. Our findings suggest that the effects of teacher PD and post-training components may vary by teacher but not student characteristics. Specifically, PD at times has small, positive and marginally significant impacts on the achievement levels of students taught by less qualified teachers. On the flip side, PD has larger, negative and significant effects on the achievement levels of students taught by more qualified teachers. In other words, PD may slightly help the least qualified teachers, but for more qualified teachers, the net effect of being out of the classroom more is ultimately negative. This effect remains after adjusting for multiple hypothesis testing.

We consider and rule out implementation failure and the substitution of teaching time as reasons for the lack of positive effects. The training program that we

evaluate was designed according to national government guidelines and was implemented with a high degree of fidelity. As such, the lack of observed effects likely derives from the program itself being ineffective, as opposed to poor implementation. Another concern is that treated teachers were not in their classrooms during the 15-day in-person training. Students were taught by a replacement teacher, however, the opportunity costs of lost time with the treated teacher could theoretically be off-set by learning gains in the long-run, which a single year of data fail to capture. Nonetheless, in the case of China's NTTP, we fail to find significantly positive average treatment effects (or treatment effects of substantive magnitude) on an extensive list of measures of teacher or student knowledge, attitudes, or behavior. Many of these intermediate outcomes are, in theory, necessary for teachers to improve student learning and, moreover, should be highly sensitive to the teacher training program in the short run. As such, the lack of change in intermediate teacher or student outcomes suggests that it is unlikely that the program, on average, will have longer-run effects on student learning.

Taken together, our findings present a cautionary tale about the ability of large-scale teacher PD programs to improve teaching and learning in developing countries. The findings are particularly concerning given the billions of dollars and teacher hours that are invested in PD programs each year. For example, between 2012 and 2017, India's national government allocated 1.2 billion USD to teacher PD programs (Government of India 2011). In Mexico, teachers spend an average of 23 days in teacher PD each year (OECD 2009). Eighty-eight percent of teachers that participated in the Teaching and Learning International Survey (TALIS) reported engaging in some PD in the last year (OECD 2014), while 63 percent of World Bank Education projects between 2000 and 2012 included PD to support teachers (Popova, Evans and Arancibia 2016). The high cost of these programs

combined with their potentially low returns should impel policymakers to re-evaluate how they invest in teacher training.

I. Experimental Design and Data

A. Intervention

China's government has invested heavily in teacher PD programs. In particular, since 2010, the government has invested more than one billion US dollars in its flagship teacher PD program—the National Teacher Training Program (MOE 2010).[4] The NTTP is the direct product of China's "2010 National Training Policy for Primary and Secondary School Teachers" which itself came from a special request to "strengthen the teacher workforce, especially among rural teachers" from the 17th Plenary Meeting of the Communist Party of China (MOE and MOF 2010).[5] The policy not only created the teacher PD program at the national level (the NTTP), but also "sister" programs throughout the country at the provincial, city, and county levels (Gong 2015; Li and Wang 2017). The NTTP was touted as the program that the sister programs should emulate (Gong 2015; Li and Wang 2017).[6] As such, the NTTP has been one of China's primary and most prominent means to improve teacher quality since 2010.

[4] Beyond 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 higher expenditures per teacher and greater prestige for participation than these local teacher PD programs.

[5] Roughly three-fourths of China's school-aged population comes from rural areas (NBS 2010), yet students in rural areas are falling far behind their urban peers on key educational outcomes. For example, while the vast majority of urban children finish high school, only 37 percent of rural children do (Shi et al. 2015). The achievement levels of rural students are also significantly lower than that of their urban peers (Loyalka et al. 2017).

[6] In the words of the 17th Plenary Congress, the NTTP was the central government's move to "send charcoal in snowy weather" (雪中送炭) or provide local governments with help in their time of need.

The amount of funding allocated towards NTTP is also significant. The direct cost of NTTP is approximately 340 million USD per annum. The direct cost of training per teacher (which include trainer salaries, the use of training facilities, equipment and materials, as well as teacher travel, room, and board) is 203 USD. The total direct cost is equivalent to approximately 19 percent of the central government's funds towards PreK-12 education. Furthermore, according to our calculations, the indirect costs of NTTP per teacher – which include the opportunity costs of teacher time from participating in both the onsite and online training – are 952 USD, making the indirect cost of the NTTP almost 5 times greater than the direct cost.

One of the major goals of the PD program is to raise teacher quality in rural regions so as to help reduce the urban-rural gap in educational outcomes (MOE 2010). Another goal is to develop a "backbone" of rural teachers that will improve the quality of colleagues who teach in the same schools (MOE 2010). In this study, we examine the impact of this flagship PD program and its two associated post-training interventions: post-training follow-up and a post-training evaluation. We describe the PD program and its associated post-training interventions in detail immediately below.[7] More details on the theoretical underpinnings of the program, a breakdown of the time spent on different content areas, and a comparison of the NTTP with PD programs in other developing countries can be found in Online Appendix A.

[7] As a third party evaluator, we had a limited role in the program. Although we conducted the randomized evaluation on behalf of the government, we had no input, influence or control over the NTTP content or sessions. We did, however, work with training providers to ensure that the post-training interventions (in addition to and after the NTTP) had a standard design and were conducted with a high degree of consistency and fidelity. Specifically, we helped the training providers develop standardized protocols for contacting teachers (for the post-training follow-up intervention) and for evaluating teacher performance (for the post-training evaluation intervention). Our enumerators also worked with the training providers to make sure the post-training interventions were implemented consistently. More specifically, for the follow-up treatment, our enumerators assisted providers in contacting the treated teachers at regular, pre-determined intervals (from the provider offices). For the evaluation treatment, our enumerators acted as official representatives of the training providers and helped score teacher performance according to the protocol.

Treatment 1: Professional Development.—The PD program, which was conducted during the academic year, focused on improving mathematics teaching in junior high schools. It consisted of two parts: (a) *in-person PD*; and (b) *supplemental online PD*.

In regards to the in-person PD, teachers participated in a 15-day onsite training at a centralized location in November 2015. Most in-service training happens during the school year, although regulations do not require this. While teachers were taken out of the classroom, the vast majority of treated students were taught mathematics by a replacement teacher for those 15 days.[8] The first two days consisted of an opening ceremony as well as an introduction and orientation to the PD program. The next 13 days consisted of morning and afternoon PD sessions of about 3 hours each. According to an analysis of syllabi and materials and daily observations of the PD sessions (conducted by our survey enumerators), the content of the PD sessions largely followed guidelines set by the Ministry of Education. The guidelines asked providers to focus on improving teacher math knowledge, pedagogy, ethics, personal growth, and classroom management strategies.

Expert trainers from university schools of education, local government bureaus of education, and math teachers from junior high schools led the PD sessions. The trainers had flexibility in deciding the format and style of the training. They were, for example, free to choose the manner in which to engage with trainees, increase trainee participation, and provide opportunities for trainee practice.

After finishing the in-person PD, trainees were able to access the online PD program. Trainees were told they could log on to an online platform at any time to peruse extra PD materials (additional slide presentations, videos, and references to

[8] In regards to the practice of taking teachers out of the classroom for training, the NTTP is similar to many PD programs in developing countries. In the TALIS sample, 50.6 percent of lower secondary education teachers across 34 countries – including those from low- and middle-income regions – report that PD conflicts with their work schedule, acting as a barrier to them receiving training (OECD 2014).

other resources). Trainees could also use the online platform to communicate with trainers and other trainees, especially to share teaching resources and discuss the application of the PD content to their classrooms. Finally, trainees were asked to turn in three short essay assignments through the online platform: (a) a brief bio; (b) a summary of one topic area covered in the PD program; and (c) an overall reflection on the PD program.

Treatment 2: Post-Training Follow-up.—Policymakers in China also emphasize the importance of regular and consistent follow-up after the in-person teacher PD sessions. Follow-up was conducted with training participants through mobile text messages and phone calls. Trainees were sent two types of messages. The first type alerted them to the existence of new, supplementary materials/assignments on the online platform. The second type provided progress reports about how much trainees had been using the online platform, the tasks they still needed to fulfill, and further encouragement to utilize the online platform. Taken together, trainees in the post-training follow-up treatment arm received about 3 messages per month. Trainees were asked to confirm the receipt of the text messages and reply with comments and questions if desired. If a trainee failed to confirm receipt of the text message within 24 hours, the trainee was called to confirm he or she had received the message.

Treatment 3: Post-Training Evaluation.—Immediately after finishing the in-service PD, trainees in the post-training evaluation treatment condition were informed that they would have to participate in an in-person evaluation, with a lesson plan and interview component, that was to be conducted at their school in two months (in January 2016, just before the midline survey). As part of the evaluation, trainees would be asked to prepare and give a 20 minute lesson plan about how they would teach students a particular math topic of their choice. The

lesson was to reflect what trainees learned from the in-person PD. Teachers would then field 5-10 minutes of questions from invitees: the school principal, other math teachers in the school, and two trained evaluators. Teachers were told that if they received a low score on the evaluation, they would not receive a completion certificate for the PD program.[9]

The evaluations were conducted according to a standardized rubric. Trainees' performance was graded separately by two evaluators, and they received points for lesson content, pedagogy, and style of delivery, especially to the degree they reflected what was learned in the on-site training. The average of the two evaluators' assessments was taken and given as feedback to the trainee.

B. Sample

Our experimental study evaluates the actual NTTP. Because of the costs associated with implementing the program, only a small percentage of teachers are selected to participate in the in-person training provided by the NTTP.[10] As it was expanding the program to reach more rural schools and teachers, the government agreed to a randomized evaluation of the NTTP. Policymakers gave us a list of the 300 rural junior high schools from 94 (out of 159) counties across a large province that were slated to participate in the NTTP. [11] Within each of the 300 schools, one grade 7-9 math teacher was selected according to a standard process: each school

[9] Teachers are incentivized to earn a certificate of completion as it is weighed in promotion decisions. Opportunities for promotion, in turn, have been found to have positive effects on teacher effort and student achievement (Karachiwalla and Park 2017).

[10] In the province where the evaluation took place, for example, between 2010 and 2016, only 1.6 percent of primary and lower secondary school teachers participated in the in-person training provided by the NTTP. The NTTP does have an online version which reaches a much wider range of teachers (approximately half of rural teachers in the province between 2010 and 2016). We chose to evaluate the in-person version of the NTTP, however, because it is more resource-intensive (and should be more impactful) and is more typical of government teacher training programs across the world.

[11] Rural schools were chosen in light of the National Teacher Training Program's focus on raising teacher quality in rural regions (MOE 2010), where school completion rates and student achievement levels are significantly lower than in urban areas (Shi et al. 2015; Loyalka et al. 2017).

nominated one teacher and this nomination was approved by the local education bureau. Selected teachers that were randomized to the treatment arms participated in the NTTP at the start of 2016. Selected teachers that were randomized to the control arm were told they would participate in the NTTP at the start of 2017. Thus, the teachers that participated in the randomized evaluation were similar to teachers that are naturally enrolled in the NTTP (absent the experiment).

We surveyed one class of students taught by each of these "primary sample" teachers. If the primary sample teacher taught more than one class of students, we randomly selected one class to be enrolled in the survey. Altogether, this primary sample consisted of 300 teachers (of which 121 teachers taught grade 7; 109 teachers taught grade 8; and 70 teachers taught grade 9) and 16,661 students.

To measure potential spillover effects from the teacher PD program, we also sampled an additional grade 7-9 math teacher and corresponding class of students within each of the 300 sample schools. Since many of the schools only had one math teacher per grade, the spillover math teacher and class were chosen from a different grade. In particular, if the primary sample teacher in a particular school was in grade 7, we randomly sampled an additional teacher and one of their classes from grade 8; if the primary sample teacher was in grade 8, we randomly sampled an additional teacher and one of their classes from grade 7; if the primary sample teacher was in grade 9, we randomly sampled an additional teacher and one of their classes from grade 7.[12] If the secondary sample teacher taught more than one class of students, we randomly selected one class to be enrolled in the survey. Altogether, this yielded an overall sample of 600 junior high math teachers and 33,580 students selected to participate in the study.

[12] Since most rural junior high schools have only one math teacher per grade, and these math teachers frequently meet together in professional learning communities at the school-level (Sargent 2015), spillovers would likely occur across any of the three grades.

C. Randomization and stratification

To estimate the impact of teacher PD and post-training interventions, we conducted a two-stage cluster-randomized trial (Figure 1). In the first stage, the 300 schools in the study were randomized, within six different blocks, to one of three treatment conditions: control or "no teacher PD" (treatment arm A in Figure 1); "teacher PD only" (treatment arm B in Figure 1); and "teacher PD plus follow-up" (treatment arm C in Figure 1).[13,14] Schools were equally distributed across treatment arms, with 100 schools in each arm. Randomly assigning teachers in this way allows us not only to evaluate the overall impact of PD, but also whether teacher PD is effective (and more effective) when it provides trainees with post-training follow-up.

In second stage randomization, half of the schools in treatment arms B and C were randomized to receive either a post-training evaluation (treatment arm X in Figure 1) or not (treatment arm Y in Figure 1). The randomization procedure ensured that the 100 schools in each of the original treatment conditions B and C had an equal probability of being assigned to one of the two post-training evaluation treatment conditions.[15]

The program intervention was characterized by a high degree of compliance. By and large, teachers that were randomly assigned to various treatments participated

[13] We randomized schools within blocks to increase statistical power. The blocks were defined by grade (grade 7, 8 or 9) and which of two agencies implemented the NTTP (yielding six blocks in total). Provincial governments in China are required to choose a small number of agencies to implement the NTTP. Agencies are chosen through a formal and rigorous bidding process. The agencies that are chosen to implement the training are from leading schools of education at the top universities within the province. We take this randomization procedure into account in our analysis by controlling for block fixed effects (Bruhn and McKenzie 2009).

[14] When the number of schools in a block was not divisible by three, we randomly allocated the remainder schools to one of the three treatment conditions. Taken together, the blocking plus randomization procedure ensured that the 300 schools in our sample had an equal probability of being assigned to one of the three treatment conditions within each block.

[15] Power calculations conducted using Optimal Design (Spybrook et al. 2009) indicate that with a pre-test intraclass correlation coefficient=0.16, R^2=0.48, beta=0.8, at least 40 students per class, and 100 schools per treatment arm, we have the power to detect a minimum detectable effect size of 0.13 SDs at the 5 percent significance level and 0.11 SDs at the 10 percent level.

in those treatments while control teachers did not receive any treatment. Specifically, 91 percent of teachers in the PD treatment group participated in PD, 98 percent of teachers in the Follow-up treatment responded to the follow-ups, and 87 percent of teachers in the Evaluation treatment completed the evaluation.

[Insert Figure 1 Here]

D. Data collection

Data collection took place in four stages: (a) administrative data collection to inform the randomization; (b) a baseline survey in October 2015; (c) a midline survey in January 2016; and (d) an endline survey in May 2016.

Administrative Data.—In the first stage, at the beginning of the academic year in October 2015, we obtained administrative data on teacher and school characteristics. Specifically, we obtained data on teacher gender, age, education level, ranking, years of experience, whether the teacher was a homeroom teacher or held an administrative position, and the number of math students they taught. We also obtained data on whether the school was rural or urban and on school size.

Student Surveys.—We collected detailed survey data on students. In the baseline survey, we asked students about their basic background characteristics: age, gender, parental education levels, and possession of household assets. During the baseline, midline, and endine surveys, we also asked students about their exposure to various teaching behaviors (including teacher care, classroom management, and instructional practices), their attitudes about math (math anxiety math self-concept, instrumental motivation for math, and intrinsic motivation for math)[16], and how

[16] We measured student attitudes towards math and teaching practices using standard scales extant in the education literature. We constructed summary indices from these scales using the GLS weighting procedure described in Anderson (2008).

much time they spent studying math each week.

Teacher Surveys.—We collected detailed data from teachers as well. In the baseline, we asked teachers to report their gender, age, years of experience, educational level (whether they went to college, whether they obtained a degree in math), rank, whether they were a homeroom teacher or not, class size, and residential (rural or urban) permit status.

During the baseline, midline and endine surveys, we collected data on a wide range of teacher attitudes and beliefs. These included teachers' intrinsic and prosocial motivation, their beliefs about the nature of math (the degree to which it is a series of rules and procedures; the degree to which it is a process of inquiry) and math teaching and learning (the degree to which math teaching should be directed; the degree to which math teaching should be active; as well as the degree to which students' math abilities are fixed). The measures capture a range of teacher beliefs that are thought to be susceptible to change and which are moderately correlated with student success in math (Clark and Peterson 1986; Fang 1996; Kagan 1992; Stipek et al. 2001; Thompson 1992). [17]

Student Standardized Mathematics Tests.—In light of the PD program's focus on improving math teaching by mathematics teachers, our primary outcome is student math achievement. Math achievement was measured at baseline, midline, and endline using 35-minute mathematics tests. The tests were grade-appropriate,

Following this procedure, we constructed a variable $\bar{s}_{ij}$ as the weighted average of k normalized outcome variables in each group (y_{ijk}), for each individual. Each dependent variable is weighted by the sum of its row entries in the inverted covariance matrix for group j, such that:

$$\bar{s}_{ij} = (1'\bar{\sum}_j^{-1}1)^{-1}(1'\bar{\sum}_j^{-1}y_{ij})$$

where 1 is a column vector of 1s, $\bar{\sum}_j^{-1}$ is the inverted covariance matrix, and y_{ij} is a column vector of all outcomes for individual i in group j. We normalize each outcome by subtracting the mean and dividing by the standard deviation, such that the summary index, $\bar{s}_{ij}$, is given in standard deviation units.

[17] We measured these teacher beliefs using internationally validated scales from Laschke and Blömeke (2013). We constructed summary indices from these scales by again using the GLS weighting procedure (Anderson 2008).

tailored to the national and provincial-level mathematics curricula. Although grade-appropriate tests may present a problem in some developing countries (since the level of student learning is already low), this was not the case in our sample schools. Our math tests were vertically scaled and showed that students, on average, made substantive gains in learning within each grade. An analysis of the test results also indicates that the tests did not suffer from floor or ceiling effects.

The tests were constructed by trained psychometricians using a multiple-stage process. Mathematics test items tests were first selected from standardized mathematics curricula for each grade (7, 8 and 9). The content validity of these test items was checked by multiple experts. The psychometric properties of the test were then validated using data from extensive pilot testing.[18]

Students took the same test at baseline and midline and a different test at endline. In the analyses, we normalized each wave of mathematics achievement scores separately using the mean and distribution in the control group. Estimated effects are thus expressed in standard deviations.

Teacher Standardized Mathematics Tests.—Teachers were given tests of math knowledge for teaching developed by researchers at the University of Michigan (Hill, Rowan and Ball 2005), at baseline, midline, and endline. These were similarly normalized. Estimated effects are thus also expressed in standard deviations.

[18] We validated the math tests used in the study using a number of steps. First, the content validity of the math tests was established by multiple math instruction experts in China. Second, after extensive piloting, we verified that the tests had good psychometric properties (Cronbach alphas of approximately 0.8, unidimensionality, and a lack of differential item functioning by gender). Third, according to the results of a separate on-going study in rural China, the scores on our math tests are moderately correlated with IQ scores from the Raven's Standard Progressive Matrices (0.47) and Wechsler Intelligence Scale for Children (0.54) tests. This falls within the range of correlations between achievement and intelligence found in the literature (Ramsay and Reynolds 2004). Finally, the distribution of math test scores (by baseline, midline and endline waves as well as by grade level) is normal and does not suffer from floor or ceiling effects. Only 1.21 percent, 2.95 percent, and 1.22 percent of students scored full marks on the baseline, midline, and endline tests respectively. Only 3.14 percent, 4.81 percent, and 5.69 percent of students scored marks that could be earned by random guessing on the baseline, midline, and endline tests respectively.

E. Balance and attrition

Online Appendix Tables B1 and B2 present balance tests on baseline teacher and student characteristics across different treatment comparison groups. Only 2 out of a total of 65 tests show statistically significant differences between treatment conditions at the 10 percent level. Another 2 out of the 65 tests show statistically significant differences at the 5 percent level. No tests are statistically different from zero at the 1 percent level. We also conduct joint tests of the significance of the baseline covariates for each of the treatment comparisons, for which p-values are all greater than 0.10. Taken together, since the number of significant differences is smaller than that expected by random chance, the randomization appears to have been successful in creating balance in baseline teacher and student characteristics across treatment conditions.[19]

We also assess the degree of differential attrition across trial arms. Overall, attrition rates were low with only 4.06 percent of students attriting by the midline and 7.85 percent attriting by the endline.[20] More importantly, cross-treatment differences in baseline student characteristics among non-attriters (Rows 1-2, 7, and 10 in Table B3) are virtually identical to cross-treatment differences in baseline student characteristics among the full baseline sample (Rows 1-2, 5, and 7 respectively in Table B2).[21] We also conduct tests for whether levels of attrition were uneven across the different treatment comparisons and find no significant differences at either midline or endline (Table B4). We therefore find no evidence of differential attrition across any of our treatment comparisons.

[19] Treatment groups (teacher PD only, teacher PD plus follow-up, and control) were also balanced in terms of the number and types of prior teacher PD opportunities they participated in (results omitted for the sake of brevity but available upon request).

[20] Students and teachers were considered to have attrited if they were not present at the midline or endline surveys.

[21] We also find no evidence of differential attrition when we look at baseline teacher characteristics (results omitted for the sake of brevity but available upon request).

F. Empirical strategy

We estimate a series of average treatment effects (ATEs). First, we compare average outcomes between (a) PD and the control group (Treatment Group B and Treatment Group A in Figure 1) and (b) PD plus post-training follow-up and the control group (Treatment Group C and Treatment Group A in Figure 1). We also estimate the ATE of the post-training follow-up intervention alone by comparing average outcomes between PD plus post-training follow-up and PD (Treatment Group B and Treatment Group C in Figure 1). Second, we compare average outcomes between PD plus post-training evaluation—conditional on whether or not the teacher also received post-training follow-up—and the control group (Treatment Group X and Treatment Group A in Figure 1). Third, we estimate the ATE of the post-training evaluation intervention alone by comparing average outcomes between PD plus post-training evaluation and PD (Treatment Group X and Treatment Group Y in Figure 1).

We estimate the ATEs using the following ordinary least squares regression model.[22]

$$Y_{ij} = \alpha_0 + \alpha_1 D_j + \mathrm{X_{ij}}\alpha + \tau_k + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the outcome of interest measured at endline for student i in school j; D_j is one or more dummies indicating the treatment assignment of school j; X_{ij} is a vector of baseline control variables, and τ_k is a set of block fixed effects. In all specifications, X_{ij} includes the baseline value of the dependent variable whenever this is available. We also estimate treatment effects with an expanded set of baseline

[22] The pre-analysis plan for the analyses was written and turned into the International Initiative for Impact Evaluation (3ie) before follow-up data were collected and before any impact analyses were run.

controls (we call these our "covariate-adjusted" regressions). For student-level outcomes, this expanded set of controls includes student age, student gender, parent educational attainment, a household asset index, class size, teacher gender, teacher age, teacher experience, teacher education level, a teacher certification dummy, a teacher major in math dummy, and teacher rank. For outcomes measured at the teacher level, student controls are omitted.

While we are primarily interested in estimating impacts on student achievement, we use the same regression specification to estimate effects on a wide range of secondary outcomes (such as student dropout, student non-cognitive outcomes, teacher knowledge, teacher attitudes, and teacher practices). By doing so, we examine potential mediators through which PD and the post-training interventions may have impacted student learning. In all cases, for dependent variables measured at the student level, we adjust standard errors for clustering at the school level using a cluster-corrected estimator. For dependent variables measured at the teacher level, we adjust standard errors using a heteroscedasticity-robust estimator. In the pre-analysis plan, we discussed adjusting standard errors for multiple hypothesis testing, however, our analyses yield p-values that are almost always above 0.1 even without adjustment.

We test for heterogeneous impacts by interacting various student and teacher baseline characteristics with the treatment indicators in equation (1). For continuous variables such as student SES, student baseline math scores, and the number of hours of PD a teacher had already accumulated prior to the study – we are particularly interested in how the effects of PD vary across the distribution of this characteristic. In these cases, we create dummy variables that capture the tercile of each distribution in which a student falls. That is, we create two new dummy variables from the continuous baseline variable. The first binary variable takes a value of 1 if the value of the continuous variable is in the top tercile, and a value of 0 otherwise. The second dummy variable takes a value of 1 if the value of the

continuous variable is in the middle tercile, and a value of 0 otherwise. These dummies are then included in the estimation procedure described above.

II. Results

Overall, none of the modalities of teacher PD has a significant impact on student achievement (Tables 1 and 2). Since the results are substantively the same whether we examine program impacts on midline or endline achievement, and with or without adjusting for covariates, we focus our discussion here on the endline results (Table 2) that adjust for covariates. Specifically, the impact of PD versus the control group is -0.006 SDs and is insignificant at the 10 percent level (Panel A, Row 1, Column 2). The estimated effect of PD plus Follow-up versus the control group is also nearly zero (0.005 SDs) and insignificant at the 10 percent level (Panel A, Row 2, Column 2). Providing teachers with PD plus Evaluation—conditional on also receiving post-training follow-up—further fails to improve student achievement relative to the control group (0.011 SDs and insignificant at the 10 percent level—Panel C, Row 8, Column 6). In fact, the upper limits of the 95 percent confidence intervals for each of the above comparisons range from 0.061 to 0.074 SDs respectively, meaning that we can convincingly rule out sizeable positive impacts.

Even the NTTP's upper bound treatment versus control estimate of 0.061 SDs is small relative to the average impact of successful interventions in developing countries. For example, interventions providing instructional inputs have been found to increase test scores significantly by an average of between 0.08 (materials) to 0.15 (computers and technology) SDs (McEwan 2015). The upper bound estimate is especially small in terms of cost-effectiveness. At the upper bound, teacher PD increases student learning by 0.27 SDs per $100, which is lower than all but two of 15 education interventions in a J-PAL (2014) study for which cost-

effectiveness data are reported.[23] The cost-effectiveness of teacher PD is also substantially lower than other educational interventions in China including teacher incentives (1.40 SDs per $100—Loyalka et al. 2016) and computer-assisted learning (4.13 SDs per $100—Mo et al. 2014).

We also find no effect of individual program components. The difference in average student achievement between PD with Follow-up versus PD only is 0.012 SDs (p-value = 0.749—Panel A, Rows 3-4, Column 2), indicating that Follow-up has no additional effect beyond PD. Similarly, PD plus Evaluation has a small, insignificant effect of 0.031 SDs relative to PD only (Panel B, Row 6, Column 4), indicating that Evaluation has no additional effect beyond PD. The small point estimates in each of these cases lie within tight 95 percent confidence intervals, once again ruling out sizeable positive impacts.

[Insert Table 1 Here]

[Insert Table 2 Here]

We also find that PD and post-training components have no impacts on a wide range of secondary student outcomes (Table 3). Neither PD only nor PD plus Follow-up has a significant impact on student dropout, math anxiety, intrinsic or instrumental motivation for math, or the amount of time spent on math (Table 3, Panel A, Rows 1-2, Columns 1-9). PD plus Evaluation also has no significant impact on any of these secondary student outcomes relative to the control group (Table 3, Panel C, Row 8, Columns 1-9). Isolating the effects of individual program components, we find no positive effect of Follow-up beyond PD (Table 3, Panel A,

[23] The two exceptions are one program which has a significant negative impact on student learning, and one conditional cash transfer program, which targets many outcomes in addition to student learning and is thus costlier than many pure education interventions. Meanwhile, the cost-effectiveness of the remaining programs in this sample with significant positive effects ranges from 1.18 to 118.34 additional SDs per $100, so even the least cost-effective of these programs is orders of magnitude more cost-effective than the upper bound cost-effectiveness of the PD program.

Rows 3-4, Columns 1-9) or of Evaluation beyond PD (Table 3, Panel B, Row 5, Columns 1-9). If anything, the addition of Evaluation to PD may slightly worsen self-concept and intrinsic motivation while increasing anxiety.

[Insert Table 3 Here]

The lack of positive effects on student outcomes is mirrored by the lack of impacts on (student-reported) teaching behaviors in the classroom (Table 4). According to our covariate-adjusted effect estimates, PD alone has an insignificant effect on all measured aspects of teacher behavior – practice, care, management, and communication (Panel A, Rows 1-2, Columns 1-4).[24] Similarly, none of the individual PD components have significant effects on any measures of teacher behavior.

[Insert Table 4 Here]

Having found no positive effects of PD and post-training components on teacher behaviors, we next examine whether they have any impact on teacher knowledge, attitudes, and beliefs. These may be important channels through which PD ultimately affects student achievement in the short or longer-term. For example, teacher beliefs about the nature of math teaching and learning are thought to be both susceptible to change and important for student success in math (Clark and Peterson 1986; Fang 1996; Kagan 1992; Stipek et al. 2001; Thompson 1992). Altogether, we find few effects of PD and post-training components on these outcomes, however, as reflected in the individual results in Table 5. Furthermore, none of the

[24] These estimates are again substantively the same as the estimates at midline (not shown for the sake of brevity but available upon request). The covariate-adjusted estimates are also similar to the covariate-unadjusted estimates (both at midline and endline) with the exception that, when compared with the control group, the coefficients on PD plus Follow-up suggest a slight deterioration in teacher practice and care (each significant at the 10 percent level) and the coefficients on PD plus Evaluation suggest a slight deterioration in teacher care (significant at the 10 percent level). These results lose significance once we adjust for multiple hypothesis testing (as specified in the pre-analysis plan), however.

results remain statistically significant at the 10 percent level after adjusting p-values for multiple hypothesis testing (as specified in the pre-analysis plan).

Given that PD and post-training components have no impacts on the outcomes of students whose teachers receive them, or on these teachers' behaviors, knowledge, attitudes or beliefs, we would not expect them to produce effects on students whose teachers did not receive PD. Indeed, we essentially find no effect of any type of PD treatment on the achievement levels of students in spillover classes (Table B5), or on the vast majority of secondary student and teacher outcomes for this sample (results not shown for the sake of brevity). While there is a slight positive impact of 0.070 SDs of PD plus Evaluation relative to PD alone, the effect is only significant at the 10 percent level and only in the analysis that does not adjust for baseline covariates.

[Insert Table 5 Here]

We finally examine whether teacher PD had differential effects on students' achievement depending on their background and that of their teachers (Table 6). We find that effects do not vary significantly by a student's household wealth (Column 1), baseline achievement level (Column 2), or the amount of training their teacher previously received (Column 3).[25] We do, however, find heterogeneous effects by teacher qualifications (Table 7). Namely, PD significantly decreases scores among students whose teachers had a college degree relative to those whose did not (-0.203 SDs). When PD is combined with Follow-up, the latter effect is even stronger (-0.312 SDs). The PD plus Follow-up also has a significant negative impact on the scores of students whose teachers majored in math relative to those whose did not (-0.143 SDs). Providing teachers with PD plus Evaluation—

[25] We also find no significant heterogeneous effects by student gender (results omitted for the sake of brevity but available upon request).

conditional on also receiving post-training Follow-up— also leads to a significant decrease in achievement for students whose teachers have college degrees relative to those whose do not (-0.254 SDs).

[Insert Table 6 Here]

[Insert Table 7 Here]

Treatment effect estimates for teacher qualification subgroups (as opposed to heterogeneous treatment effect estimates between subgroups) are similar, even after adjusting p-values for multiple hypothesis testing.[26] In particular, we find that relative to the control group: (a) PD plus follow-up and PD (only) have negative effects on the achievement of students whose teachers went to four year college (-0.215 and -0.147 SDs respectively, both significant at the 5 percent level, adjusted p-values of 0.003 and 0.035); (b) PD plus follow-up has small, positive effects on the achievement of students whose teachers did not go to four year college (0.097 SDs, significant at the 5 percent level, adjusted p-value = 0.047); (c) PD plus evaluation has negative effects on the achievement of students whose teachers went to four year college (-0.167 SDs, significant at the 5 percent level, adjusted p-value = 0.033) and smaller, positive effects on the achievement of students whose teachers did not go to four year college (0.087 SDs, significant at the 5 percent level, adjusted p-value = 0.004).

Taken together, these exploratory findings suggest that teacher PD has moderately sized, negative effects among more qualified teachers and, at best, only slight positive effects among less qualified teachers. This is likely because the

[26] Although we stated in our pre-analysis plan that we would not adjust p-values for multiple hypothesis testing when analyzing impacts on subgroups (since we treat the analyses as exploratory), we do provide adjusted the p-value estimates here for the impacts of different combinations of PD and post-training intervention components relative to the control group for six different subgroups (female teachers, male teachers, teachers with and without a college degree, and teachers with and without a math major). Results are not shown in a separate table for the sake of brevity but are available upon request.

teacher PD program causes teachers to substitute their own time away from teaching. If more qualified teachers were originally helping students learn (more than potentially less-qualified replacement teachers), while less qualified teachers were perhaps not contributing to learning, then the loss in teaching time of qualified teachers would have negative consequences for student learning.

A. Why does PD not work?

The above results show that student achievement, psychological traits related to achievement, effort, and dropout are not affected by teacher PD. More proximally, teacher knowledge, attitudes, and behaviors are not affected by PD either. How do we explain the lack of significant impacts on such a wide range of student and teacher outcomes? To explore this question further, we examine several hypothesized mechanisms which, in the causal chain, precede changes in teacher knowledge, attitudes and behavior (as well as, of course, student outcomes). These hypothesized mechanisms include (a) the degree to which trainees participated in the PD sessions and post-training interventions; (b) the accessibility and relevance of PD content; (c) whether PD was delivered in an impactful way; and (d) whether teachers had adequate resources and were free from constraints to implement what they learned from the PD sessions. We use several additional sources of data to examine whether these mechanisms are likely to be responsible for the lack of significant impacts: (i) observations of participant behavior in the in-person PD sessions, online PD sessions, and evaluations; (ii) syllabi and course content of the in-person and online PD sessions; and (iii) in-depth interviews with 40 teachers that participated in the various PD treatment conditions.[27]

[27] Specifically, we randomly selected and interviewed 10 teachers who participated in the PD program and who received no post-training interventions, 10 teachers who participated in the PD program and received Follow-up but no Evaluation, 10 teachers who participated in the PD program and received an Evaluation but no Follow-up, and 10 teachers who

Trainee participation was high. According to our records, daily attendance for the on-site PD sessions was 93 percent.[28] In addition, daily observations from our enumerators revealed that, throughout the on-site PD sessions, trainees exhibited relatively high levels of attention and interest, as well as positive attitudes to learn.[29] Teachers further watched an average of 17 hours of video lectures, commented in chat rooms an average of 24 times, and received an average grade of 95.8 out of 100 points on the three brief assignments associated with the online PD. Finally, the 9 out of 10 teachers that were assigned to the Evaluation treatment condition delivered their prepared lesson plan, passed the assessment criteria, and received evaluative feedback.

Although trainee participation was high, the content of the program interventions was not particularly accessible or relevant. An analysis of the course syllabi and materials revealed that approximately 52 percent of the materials were "theoretical" with little application to the real world. Details of this analysis are provided in Online Appendix A. Moreover, in interviews, teachers stated that the majority of the content of the on-site and online PD was difficult and unrealistic. Approximately 88 percent of the teachers stated that they wished the content were more practical, as opposed to theoretical. As one teacher stated "we were taught 24 different teaching strategies, none of which we felt we could apply in practice." Teachers further noted that new techniques, such as having students work together in small groups or using assessment data to improve pedagogy, were only introduced as abstract concepts. Teachers felt ill-equipped to apply these abstract concepts within their classrooms or to share them with their fellow teachers.

participated in the PD program and received both Follow-up and an Evaluation. Teachers were interviewed after the endline survey.

[28] Approximately 92 percent of trainees attended more than 13 out of 15 days of on-site PD.

[29] Enumerators used a detailed protocol to score teacher attention, interest, and attitudes. On average, teachers received 4.3 out of 5 points in each of these three areas.

Further exacerbating the ability of trainees to absorb and learn from the PD sessions was the markedly passive and rote delivery of PD content. According to our enumerator's daily logs, trainers used the vast majority of the on-site PD sessions to lecture. Only in a minority of cases did trainers leave a few minutes at the end of the session for questions and answers. Many of the interviewed teachers noted that the training was not impactful precisely because there was little time for dialogue and interaction with the trainers. The online PD sessions, largely consisting of video lectures, were similarly passive in nature. Trainees reported that they were busy with their daily duties as teachers and only gave cursory attention to the online content, which they often let run in the background.

Finally, some teachers reported being constrained in trying to apply the practices they did learn from PD in their classrooms. Several of the teachers that we interviewed stated that new technologies were introduced during the PD sessions (such as the use of a multimedia graphing tool) but that they had no access to those technologies in their schools. Some teachers also complained that the heavy and fast-paced curricula of junior high schools left little room for new types of teaching practices or classroom management styles. Others noted that the large degree of heterogeneity in student ability in their classrooms also prohibited them from applying new teaching techniques.[30]

III. Conclusion

Governments spend billions of dollars and billions of hours of teacher time on teacher PD programs each year, yet the effectiveness of these programs is not well

[30] In light of the comments from the teacher interviews, we tested the hypothesis that training was too difficult to implement because student ability was so diverse. To test this hypothesis, we ran heterogeneous effects analyses by interacting treatment with the within-class coefficient of variation, as well as the within-class SD, in baseline student test scores as ex-ante measures of dispersion. The magnitude of the interaction effects is close to zero and not statistically significant (results omitted for the sake of brevity). We thus find no evidence that training was too hard to implement because student ability was too diverse.

understood. The results of this study indicate that neither teacher PD alone nor PD combined with follow-up and/or evaluation have any significant impacts on student achievement, dropout, or subject-specific psychological factors. PD also has no impact on teacher knowledge, attitudes, or teaching practices that might lead to impacts on students in the longer term. Our findings on the impacts of teacher PD are consistent with a more general literature that often finds no impacts from long-term, intensive job training on employment and wages (Card, Kluve, and Weber 2010; Ibarrarán and Rosas Shady 2008; U.S. Department of Labor 2014). Our findings do suggest some heterogeneous effects, however, with PD and its post-training components having small, positive effects on the achievement of students taught by less qualified teachers, and larger, negative effects on the achievement of students of more qualified teachers.

Our study makes four major contributions to the literature. First, to the best of our knowledge, this is the first large-scale randomized evaluation of teacher PD in K-12 schooling in a developing country. Second, this is one of the first evaluations of post-training interventions that hypothetically strengthen teacher PD. Third, unlike most studies, we conduct a thorough analysis of the causal chain, suggesting reasons for the lack of impacts. Fourth and finally, this is the first experimental evaluation of a government-sponsored teacher PD policy in a developing country. Most experimental evaluations of teacher PD programs are efficacy studies that are implemented with a high degree of fidelity (Yoon et al. 2007; Fryer 2016), usually through researcher-run pilots. In contrast, our study evaluates the impacts of a teacher PD program that was sponsored under a more realistic, policy-relevant context.

Our study has important implications for education policymakers. Teacher PD has no effects even in China, which among developing countries has a relatively well-organized education system with ample resources to fund and manage PD programs. Policymakers in China are highly selective in choosing PD providers, to

whom they give clear guidelines on designing PD content. The duration of the on-site PD program is substantial as are the online resources it provides. Teachers attend the PD sessions and make use of the extensive online resources. Even in this amenable context, however, PD has no impact. At best, heterogeneous responses to treatment from different teachers suggest that teacher PD programs may need to move beyond one-size-fits-all approaches. Policymakers in countries with fewer resources may thus wish to proceed cautiously in promoting PD programs.

Our study also highlights the importance of rigorous policy impact evaluation. Policymakers in China and elsewhere have relied on teachers' high (self-reported) satisfaction ratings to conclude that PD programs are effective. Indeed, in our study teachers report an average satisfaction level of 4.5 out of 5. However, a closer probing of our interview data shows that this satisfaction is driven by the material conditions of the training site and the way teachers are treated by the PD provider, and not by any perceived improvements in their teaching and ultimately student learning due to the PD. Reliance on such misleading data for evaluating PD can lead to misinformed policy decisions.

Our study's examination of the effectiveness of post-training interventions, such as follow-up and evaluation, offer additional insights for policymakers. Lower cost solutions such as follow-up reminders via text message or phone call and one-off evaluations may do little to increase the effectiveness of PD when the content and delivery of PD is overly theoretical. Instead, more human resource intensive follow-up (mentoring visits) and evaluation (formative assessment) may be more effective (Guskey 2002; Hobson et al. 2009; Popova, Evans and Arancibia 2016). Our study, of course, does not speak to the effectiveness of these types of human resource intensive interventions. Furthermore, such interventions are more difficult for policymakers in developing countries to implement given their higher costs and greater demands on technical expertise and implementation capacity.

While our findings are not necessarily generalizable to other countries and contexts, this study again serves as a cautionary tale for policymakers interested in improving the quality of their teacher labor force. Given the massive emphasis and government expenditures on teacher PD, policymakers in other developing countries–with often fewer resources and organizational capacity than China–may wish to reconsider their current PD programs. This reconsideration could take three possible forms. First, governments may wish to invest efforts in rigorously evaluating the effectiveness of the content and delivery methods of their current programs. Second, given the billions of dollars spent each year on PD, policymakers may wish to consider investing in other types of PD programs that find more support in education theory and practice.[31] Likewise, they may wish to revisit decisions to introduce low-cost but potentially ineffective PD components, such as those that exploit technology as a substitute for human trainers. Finally, if the costs involved in building capacity to implement other types of PD programs are prohibitive (or if indeed these PD programs are also minimally effective), policymakers may consider diverting resources into other possible ways of improving the quality of the teaching force.

[31] For example, PD that includes detailed instructions on implementation and an even larger number of contact and support hours (Yoon et al. 2007; Fryer 2016)

REFERENCES

Anderson, Michael L. 2008. “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” *Journal of the American Statistical Association* 103 (484): 1481–1495. doi:10.1198/016214508000000841.

Ball, Arnetha F. 2000. “Preparing Teachers for Diversity: Lessons Learned from the US and South Africa.” *Teaching and Teacher Education* 16 (4): 491–509. doi:10.1016/s0742-051x(00)00007-x.

Bau, Natalie, and Jishnu Das. 2017. “The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers.” Policy Research Working Paper 8050. World Bank. doi:10.1596/1813-9450-8050.

Behrman, Jere R., Shahrukh Khan, David Ross, and Richard Sabot. 1997. "School quality and cognitive achievement production: A case study for rural Pakistan." *Economics of Education Review* 16 (2): 127-142. doi:10.1016/s0272-7757(96)00045-3.

Behrman, Jere R., David Ross, and Richard Sabot. 2008. “Improving Quality Versus Increasing the Quantity of Schooling: Estimates of Rates of Return from Rural Pakistan.” *Journal of Development Economics* 85 (1–2): 94–104. doi:10.1016/j.jdeveco.2006.07.004.

Berry, Barnett, ed. 2011. *Teaching 2030: What we must do for our students and our public schools: Now and in the future*. New York: Teachers College Press.

Bold, Tessa; Filmer, Deon; Martin, Gayle; Molina, Ezequiel; Rockmore, Christophe; Stacy, Brian; Svensson, Jakob; and Wane, Waly. 2017. “What Do Teachers Know and Do? Does It Matter?: Evidence from Primary Schools in Africa.” Policy Research Working Paper 7956. World Bank.

Braslavsky, Cecilia and Birgin, Alejandra. 1992. “Situación del magisterio argentino y aportes para el diseño de estrategias de capacitación.” Serie

Documentos e Informes de Investigación 136. *Facultad Latinoamericana de Ciencias Sociales, Programa Buenos Aires.*

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232. doi:10.1257/app.1.4.200.

Bruns, Barbara, and Javier Luque. 2015. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank. doi:10.1596/978-1-4648-0151-8.

Buhl-Wiggers, Julie, Jason Kerwin, Jeffrey Smith, and Rebecca Thornton. 2017. "The impact of teacher effectiveness on student learning in Africa." *Centre for the Study of African Economies Conference*.

Card, David, Jochen Kluve, and Andrea Weber.2010. "Active Labour Market Policy Evaluations: A Meta-Analysis." *The Economic Journal* 120 (548): F452–F477. doi:10.1111/j.1468-0297.2010.02387.x.

Carpenter, Thomas P., Elizabeth Fennema, Penelope L. Peterson, Chi-Pang Chiang, and Megan Loef. 1989. "Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study." *American Educational Research Journal* 26 (4): 499-531. doi:10.3102/00028312026004499.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–2679. doi:10.1257/aer.104.9.2633.

Clark, Christopher M., and Penelope L. Peterson. 1986. "Teachers' thought processes." In *Handbook of Research in Teaching*, edited by Merlin Wittrock, 255–296 New York: MacMillan.

Cobb, Velma L. 1999. "An International Comparison of Teacher Education." ERIC Digests ED436486.

Cochran-Smith, Marilyn, and Susan L. Lytle. 1999. "Relationships of Knowledge

and Practice: Teacher Learning in Communities." *Review of Research in Education* 24: 249. doi:10.2307/1167272.

Cohen, David K. 1990. "A Revolution in One Classroom: The Case Of Mrs. Oublier." *Educational Evaluation and Policy Analysis* 12 (3): 311-329. doi:10.2307/1164355.

Corcoran, Thomas B. 1995. "Helping Teachers Teach Well: Transforming Professional Development." CPRE Policy Brief ED388619.

Darling-Hammond, Linda, Marcella L. Bullmaster, and Velma L. Cobb. 1995. "Rethinking Teacher Leadership Through Professional Development Schools." *The Elementary School Journal* 96 (1): 87–106. doi:10.1086/461816.

Darling-Hammond, Linda, and Milbrey W. McLaughlin. 2011. "Policies that support professional development in an era of reform." *Phi delta kappan* 92 (6): 81–92. doi:10.1177/003172171109200622.

Das, Jishnu, and Tristan Zajonc. 2010. "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement." *Journal of Development Economics* 92 (2): 175–187. doi:10.1016/j.jdeveco.2009.03.004.

Desimone, Laura M. 2009. "Improving impact studies of teachers' professional development: Toward better conceptualizations and measures." *Educational researcher* 38 (3): 181-199. doi:10.3102/0013189x08331140.

Evans, David K., and Anna Popova. 2016. "Cost-effectiveness analysis in development: Accounting for local costs and noisy impacts." *World Development* 77: 262-276. doi:10.1016/j.worlddev.2015.08.020.

Fang, Zhihui. 1996. "A Review of Research on Teacher Beliefs and Practices." *Educational Research* 38 (1): 47–65. doi:10.1080/0013188960380104.

Foster, Phillip. (1977). "Educational and Social Differentiation in Less Developed Countries." *Comparative Education Review* 21(2–3): 211–29.

Fryer, Roland. 2016. "The Production of Human Capital in Developed Countries:

Evidence from 196 Randomized Field Experiments." NBER Working Paper 22130. National Bureau of Economic Research. doi:10.3386/w22130.

Ganser, Tom. 2000. "An ambitious vision of professional development for teachers." *NASSP bulletin* 84 (618): 6-12. doi:10.1177/019263650008461802.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa et al. 2008. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." NCEE 2008-4030. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Marian Eaton, Kirk Walters, Mengli Song et al. 2011. "Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation." NCEE 2011-4024. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Gersten, Russell, Mary Jo Taylor, Tran D. Keys, Eric Rolfhus, and Rebecca Newman-Gonchar. 2014. "Summary of Research on the Effectiveness of Math Professional Development Approaches." Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, U.S. Department of Education; and Regional Educational Laboratory Southeast at Florida State University.

Gong, Yanfang. 2015. "Rules, Protocols, and Characteristic—Tongliao Strengthens Efforts in Teacher Training." *Inner Mongolia Education* 22: 28. (In Chinese).

Government of India. 2011. Report of the Working Group on Teacher Education for the 12th Five Year Plan. Department of School Education and Literacy, Ministry of Human Resource Development, Government of India.

Guskey, Thomas R. 1995. "Professional development in education: In search of the

optimal mix.” In *Professional development in education: New paradigms and practices*, edited by Guskey, Thomas R., and Michael Huberman, New York: Teachers College Press.

Guskey, Thomas R. 2002. "Professional development and teacher change." *Teachers and Teaching* 8 (3): 381-391.

Hanushek, Eric A, and Steven G Rivkin. 2010. “Generalizations About Using Value-Added Measures of Teacher Quality.” *American Economic Review* 100 (2): 267–271. doi:10.1257/aer.100.2.267.

Hiebert, James, and Douglas 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, Heather C., Brian Rowan, and Deborah Loewenberg Ball. 2005. “Effects of teachers’ mathematical knowledge for teaching on student achievement.” *American educational research journal*, 42 (2): 371-406. doi:10.3102/00028312042002371.

Hobson, Andrew J., Patricia Ashby, Angi Malderez, and Peter D. Tomlinson. 2009. "Mentoring beginning teachers: What we know and what we don't." *Teaching and Teacher Education* 25 (1): 207-216. doi:10.1016/j.tate.2008.09.001.

Ibarrarán, Pablo, and David Rosas Shady. 2009. "Evaluating the impact of job training programmes in Latin America: evidence from IDB funded operations." *Journal of Development Effectiveness* 1 (2): 195-216. doi:10.1080/19439340902918094.

J-PAL. 2014. “Student learning and student attendance cost-effectiveness analysis data.” Jameel Poverty Action Lab. www.povertyactionlab.org/doc/cea-data-full-workbook (accessed August 2014).

Kagan, Dona M.1992. "Implication of research on teacher belief." *Educational psychologist* 27 (1): 65-90.

Karachiwalla, Naureen, and Albert Park. 2017. "Promotion incentives in the public

sector: evidence from Chinese schools." *Journal of Public Economics* 146: 109-128.

Kennedy, Mary. 1998. "Form and substance in in-service teacher education" Research Monograph No. 13.

Laschke, Christin, and Sigrid Blömeke, eds. 2014. *Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M 2008). Dokumentation der Erhebungsinstrumente*. Waxmann Verlag, 2014.

Li, Jinyu, and Wang, Jian. 2017. "The Innovative Contribution of the National Teacher Training Program to the Teacher Training in China." *Research on Teacher Development* (2): 1-9. (In Chinese).

Lieberman, Ann. 1995. "Practices that support teacher development." *Phi delta kappan* 76 (8): 591.

Loyalka, Prashant, James Chu, Jianguo Wei, Natalie Johnson, and Joel Reniker. 2017. "Inequalities in the pathway to college in China: When do students from poor areas fall behind?" *The China Quarterly* 229: 172-194. doi:10.1017/s0305741016001594.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi. 2016. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." REAP Working Paper.

McEwan, Patrick J. 2015. "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments." *Review of Educational Research* 85, (3): 353-394. doi:10.3102/0034654314553127.

Metzler, Johannes, and Ludger 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.

Mo, Di, Linxiu Zhang, Renfu Luo, Qinghe Qu, Weiming Huang, Jiafu Wang, Yajie Qiao, Matthew Boswell, and Scott Rozelle. 2014. "Integrating Computer-

Assisted Learning into a Regular Curriculum: Evidence from a Randomised Experiment in Rural Schools in Shaanxi." *Journal of Development Effectiveness* 6 (3): 300–323. doi:10.1080/19439342.2014.911770.

MOE. 2010. Notice from the Ministry of Education and Ministry of Finance of the Implementation of the National Training Plan for Primary and Secondary Education Teachers. Ministry of Education and Ministry of Finance, Government of China.

MOE and MOF. 2010. Circular of the Ministry of Education and the Ministry of Finance on the Implementation of "National Teacher Training Program for Primary and Secondary School Teachers". Ministry of Education. http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/s4645/201212/146071.html

OECD. 2009. *Creating Effective Teaching and Learning Environments: First Results from TALIS.* Teaching and Learning International Survey. Organisation for Economic Co-Operation and Development. doi:10.1787/9789264068780-en.

OECD. 2014. *TALIS 2013 Results: An International Perspective on Teaching and Learning.* Teaching and Learning International Survey. Organisation for Economic Co-operation and Development. doi:10.1787/9789264196261-4-en.

Popova, Anna, David K. Evans, and Violeta Arancibia. 2016. "Inside in-service teacher training: What works and how do we measure it?" World Bank Policy Research Working Paper 7834.

Popova, Anna, David K. Evans, Mary E. Breeding, and Violeta Arancibia. 2018. "Global Landscape of Teacher Professional Development Programs: The Gap between Evidence and Practice." Unpublished working paper, last modified March 1, 2018, World Bank.

Peressini, Dominic, Hilda Borko, Lew Romagnano, Eric Knuth, and Christine Willis. 2004. "A conceptual framework for learning to teach secondary mathematics: A situative perspective." *Educational Studies in Mathematics* 56

(1): 67-96. doi:10.1023/b:educ.0000028398.80108.87.

Ramsay, Michael. C., and Cecil Reynolds, C. 2004. “Relations between intelligence and achievement tests.” *Comprehensive handbook of psychological assessment* 25-50. doi:10.1002/9780471726753.ch3.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94 (2): 247–252. doi:10.1257/0002828041302244.

Sargent, Tanja C. 2015. “Professional Learning Communities and the Diffusion of Pedagogical Innovation in the Chinese Education System.” *Comparative Education Review* 59 (1): 102–132. doi:10.1086/678358.

Schifter, Deborah, Susan Jo Russell, and Virginia Bastable. 1999. "Teaching to the big ideas." In *The diagnostic teacher: Constructing new approaches to professional development*, edited by Mildred Z Solomon 22-47, New York: Teachers College Press.

Shepherd, Debra. 2015. “Learn to teach, teach to learn: A within-pupil across-subject approach to estimating the impact of teacher subject knowledge on South African grade 6 performance.” Stellenbosch Economic Working Paper 01/15.

Shi, Yaojiang, Linxiu Zhang, Yue Ma, Hongmei Yi, Chengfang Liu, Natalie Johnson, James Chu, Prashant Loyalka, and Scott Rozelle. 2015. "Dropping out of rural China's secondary schools: A mixed-methods analysis." *The China Quarterly* 224: 1048-1069.

Silva, Eduardo C. 1991. “La formación docente en América Latina: un desafío que requiere respuesta.” Santiago de Chile: OREALC/UNESCO.

Spybrook, Jessaca, Stephen W. Raudenbush, Xiaofeng Liu, Richard Congdon, and Andrés Martínez. 2009. Optimal Design for Longitudinal and Multilevel Research v1.76 [Computer Software].

Stipek, Deborah J., Karen B. Givvin, Julie M. Salmon, and Valanne L. MacGyvers. 2001. "Teachers’ beliefs and practices related to mathematics

instruction." *Teaching and teacher education* 17 (2): 213-226.

Subirats, José, and Ivonne Nogales. 1989. "Maestros, escuelas, crisis educativa. Condiciones del trabajo docente en Bolivia." Santiago de Chile: OREALC/UNESCO.

Tandon, Prateek, and Tsuyoshi Fukao. 2015. "Educating the Next Generation: Improving Teacher Quality in Cambodia." Washington, DC: World Bank. doi:10.1596/978-1-4648-0417-5.

Thompson, Alba, G. 1992. "Teachers' beliefs and conceptions: A synthesis of the research." In *Handbook of research on mathematics teaching and learning*, edited by Douglas Grouws, 127–146 New York: MacMillan.

U.S. Department of Labor. 2014. "What works in job training: A synthesis of the evidence." U.S. Department of Labor, U.S. Department of Commerce, U.S. Department of Education, and U.S. Department of Health and Human Services.

Vegas, Emiliana. 2007. "Teacher Labor Markets in Developing Countries." *The Future of Children* 17 (1): 219–232. doi:10.1353/foc.2007.0011.

Villegas-Reimers, Eleonora. 1998. *The preparation of teachers in Latin America: Challenges and trends*. World Bank.

Villegas-Reimers, Eleonora. 2003. *Teacher PD: An international review of the literature*. Paris: International Institute for Educational Planning.

Westfall, P. H., and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons.

Yoon, Kwang S., Duncan, Teresa, Lee, Silvia WY, Scarloss, Beth, and Kathy L. Shapley (2007). "Reviewing the Evidence on How Teacher PD Affects Student Achievement. Issues and Answers." REL Working Paper 2007-33. *Regional Educational Laboratory Southwest*.

Yoshikawa, Hirokazu, Diana Leyva, Catherine E. Snow, Ernesto Treviño, M. Clara Barata, Christina Weiland, Celia J. Gomez, et al. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool

Classroom Quality and Child Outcomes." *Developmental Psychology 51* (3): 309–322. doi:10.1037/a0038785.

Zepeda, Sally J. 2012. *Professional development: What works* (2nd ed.). Larchmont, NY: Eye on Education.

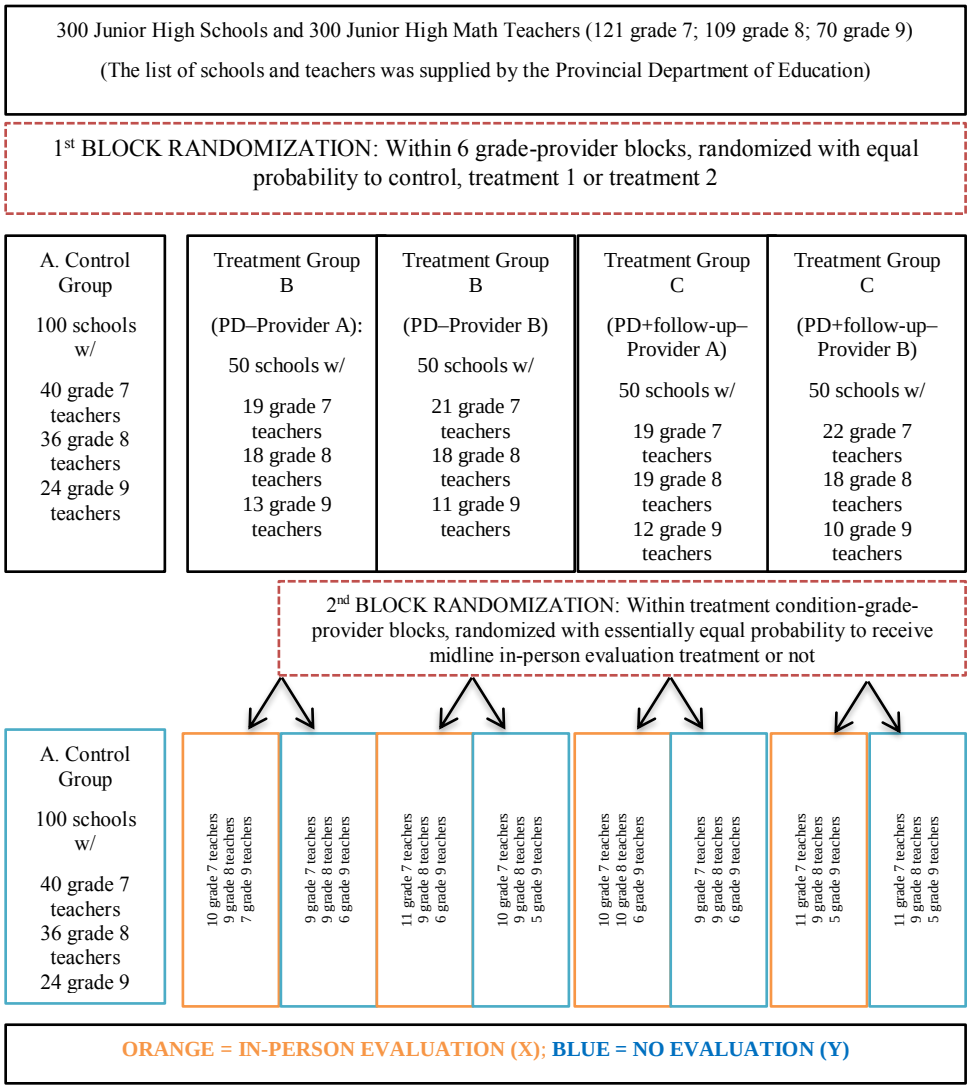

FIGURE 1. RANDOMIZATION PROCEDURE

TABLE 1 – IMPACTS ON STUDENT ACHIEVEMENT (AT MIDLINE)

		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)							
(1)	PD	-0.015	-0.035				
		(0.028)	(0.027)				
(2)	PD + Follow-up	0.000	-0.020				
		(0.031)	(0.030)				
(3)	Difference: PD + Follow-up - PD	0.015	0.015				
(4)	P-value: PD + Follow-up - PD	0.609	0.613				
(5)	Observations	15,987	15,713				
Panel B: Comparing PD + Evaluation versus PD (left-out group)							
(6)	PD + Evaluation			0.008	0.005		
				(0.029)	(0.028)		
(7)	Observations			10,725	10,483		
Panel C: Comparing PD + Evaluation versus Control (left-out group)							
(8)	PD + Evaluation					-0.003	-0.022
						(0.028)	(0.028)
(9)	Observations					10,967	10,774
(10)	Additional controls		X		X		X

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted student and teacher baseline covariates and block fixed effects. PD stands for professional development. According to the standard error estimates, none of the coefficients are statistically significant at even the 10 percent level. Of course, after adjusting p-values for multiple hypothesis testing (using the Free Step-Down Resampling Method of Westfall and Young (1993)), the estimated coefficients remain statistically insignificant at the 10 percent level.

TABLE 2 – IMPACTS ON STUDENT ACHIEVEMENT (AT ENDLINE)

		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)							
(1)	PD	0.023	-0.006				
		(0.036)	(0.034)				
(2)	PD + Follow-up	0.026	0.005				
		(0.037)	(0.035)				
(3)	Difference: PD + Follow-up - PD	0.003	0.012				
(4)	P-value: PD + Follow-up - PD	0.934	0.749				
(5)	Observations	14,838	14,599				
Panel B: Comparing PD + Evaluation versus PD (left-out group)							
(6)	PD + Evaluation			0.043	0.031		
				(0.037)	(0.034)		
(7)	Observations			9,934	9,726		
Panel C: Comparing PD + Evaluation versus Control (left-out group)							
(8)	PD + Evaluation					0.044	0.011
						(0.035)	(0.032)
(9)	Observations					10,168	10,006
(10)	Additional controls		X		X		X

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted student and teacher baseline covariates and block fixed effects. PD stands for professional development. According to the standard error estimates, none of the coefficients are statistically significant at even the 10 percent level. Of course, after adjusting p-values for multiple hypothesis testing (using the Free Step-Down Resampling Method of Westfall and Young (1993)), the estimated coefficients remain statistically insignificant at the 10 percent level.

TABLE 3 - IMPACTS ON SECONDARY STUDENT OUTCOMES (AT ENDLINE)

		Dropout (yes/no)	Math self-concept	Math anxiety	Intrinsic motivation	Instrumental motivation	Time on math
		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Comparing PD and PD + Follow-up versus Control (left-out group)							
(1)	PD	-0.002	0.041	0.007	-0.009	0.029	0.025
		(0.009)	(0.028)	(0.022)	(0.037)	(0.035)	(0.058)
(2)	PD + Follow-up	0.001	0.013	0.001	-0.018	-0.013	-0.024
		(0.009)	(0.029)	(0.024)	(0.036)	(0.034)	(0.060)
(3)	Difference: PD + Follow-up - PD	0.003	-0.029	-0.006	-0.009	-0.042	-0.049
(4)	P-value: PD + Follow-up - PD	0.757	0.285	0.790	0.792	0.162	0.406
(5)	Observations	16,305	14,475	14,442	14,533	14,548	14,323
Panel B: Comparing PD + Evaluation versus PD (left-out group)							
(6)	PD + Evaluation	-0.009	-0.054	0.054	-0.075	-0.045	0.005
		(0.008)	(0.026)	(0.023)	(0.033)	(0.029)	(0.054)
(7)	Observations	10,862	9,649	9,623	9,680	9,692	9,545
Panel C: Comparing PD + Evaluation versus Control (left-out group)							
(8)	PD + Evaluation	-0.005	-0.002	0.036	-0.046	-0.008	0.002
		(0.008)	(0.029)	(0.024)	(0.037)	(0.035)	(0.061)
(9)	Observations	11,165	9,918	9,901	9,968	9,976	9,806

Notes: Cluster-robust SEs in parentheses. Estimates adjusted for student and teacher baseline covariates and block fixed effects. PD stands for professional development. After adjusting p-values for multiple hypothesis testing using the Free Step-Down Resampling Method (Westfall and Young 1993), none of the estimated coefficients are significant at the 10 percent level. Specifically, four coefficients became statistically insignificant after adjusting p-values [Row 6, Column 2, adjusted p-value = 0.304; Row 6, Column 3, adjusted p-value = 0.189; Row 6, Column 4, adjusted p-value = 0.199].

TABLE 4 - IMPACTS ON TEACHER PRACTICE (AT ENDLINE)

		Teacher practice	Teacher care	Teacher management	Teacher communication
		(1)	(2)	(3)	(4)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)					
(1)	PD	0.043	0.028	0.008	0.051
		(0.045)	(0.046)	(0.048)	(0.048)
(2)	PD + Follow-up	-0.069	-0.060	-0.022	-0.023
		(0.046)	(0.044)	(0.047)	(0.046)
(3)	Difference: PD + Follow-up - PD	-0.111	-0.088	-0.031	-0.074
(4)	P-value: PD + Follow-up - PD	0.021	0.041	0.544	0.123
(5)	Observations	14,405	14,550	14,582	14,583
Panel B: Comparing PD + Evaluation versus PD (left-out group)					
(6)	PD + Evaluation	-0.018	-0.069	-0.000	-0.073
		(0.045)	(0.042)	(0.051)	(0.045)
(7)	Observations	9,589	9,697	9,712	9,712
Panel C: Comparing PD + Evaluation versus Control (left-out group)					
(8)	PD + Evaluation	-0.020	-0.052	0.007	-0.010
		(0.044)	(0.045)	(0.047)	(0.048)
(9)	Observations	9,872	9,970	9,995	10,002

Notes: Cluster-robust SEs in parentheses. All estimates adjusted for student and teacher baseline covariates and block fixed effects. PD stands for professional development. According to the standard error estimates, only two of the coefficients are statistically significant at even the 10 percent level (Panel A, Row 4, Columns 1 and 2). After adjusting p-values for multiple hypothesis testing (using the Free Step-Down Resampling Method of Westfall and Young (1993)), the estimated coefficients are statistically insignificant at the 10 percent level (adjusted p-values of 0.404 and 0.606 respectively).

TABLE 5 - IMPACTS ON TEACHER KNOWLEDGE AND ATTITUDES (AT ENDLINE)

		Teacher math knowledge	Teacher Intrinsic motivation	Teacher prosocial motivation	Teacher belief in directed math learning	Teacher belief in active math learning	Teacher belief that math ability is fixed	Teacher belief in math nature as rules and procedures	Teacher belief in math nature as process of inquiry
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)									
(1)	PD	0.153	0.057	-0.064	-0.263	-0.235	0.111	-0.247	-0.130
		(0.138)	(0.124)	(0.133)	(0.131)	(0.130)	(0.124)	(0.149)	(0.156)
(2)	PD + Follow-up	0.222	-0.033	-0.049	-0.145	-0.076	0.205	0.070	0.122
		(0.145)	(0.128)	(0.134)	(0.133)	(0.127)	(0.123)	(0.131)	(0.125)
(3)	Difference: PD + Follow-up - PD	0.068	-0.090	0.015	0.118	0.159	0.094	0.317	0.251
(4)	P-value: PD + Follow-up - PD	0.580	0.506	0.912	0.353	0.290	0.479	0.034	0.085
(5)	Observations	293	295	295	295	295	294	295	295
Panel B: Comparing PD + Evaluation versus PD (left-out group)									
(6)	PD + Evaluation	0.044	0.275	0.109	0.094	-0.020	-0.063	-0.067	0.048
		(0.121)	(0.124)	(0.133)	(0.121)	(0.144)	(0.120)	(0.146)	(0.145)
(7)	Observations	192	194	194	194	194	193	194	194
Panel C: Comparing PD + Evaluation versus Control (left-out group)									
(8)	PD + Evaluation	0.271	0.110	0.005	-0.215	-0.180	0.111	-0.167	-0.014
		(0.136)	(0.123)	(0.131)	(0.128)	(0.139)	(0.127)	(0.145)	(0.162)
(9)	Observations	201	202	202	202	202	202	202	202

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for teacher baseline covariates and block fixed effects. PD stands for professional development. After adjusting p-values for multiple hypothesis testing using the Free Step-Down Resampling Method (Westfall and Young 1993), none of the estimated coefficients are significant at the 10 percent level [Row 1, Column 4, adjusted p-value = 0.645; Row 1, Column 5, adjusted p-value = 0.905; Row 1, Column 7, adjusted p-value = 0.690; Row 2, Column 6, adjusted p-value = 0.818; Row 4, Column 7, adjusted p-value = 0.399; Row 4, Column 8, adjusted p-value = 0.681; Row 6, Column 2, adjusted p-value = 0.364; Row 8, Column 1, adjusted p-value = 0.497; Row 8, Column 4, adjusted p-value = 0.741].

TABLE 6 – IMPACTS ON STUDENT ACHIEVEMENT BY STUDENT AND TEACHER GROUP TERCILES (AT ENDLINE)

		Student household (1)	Student (2)	Hours teacher (3)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)				
(1)	PD	0.035	-0.019	-0.067
		(0.045)	(0.046)	(0.073)
(2)	PD + Follow-up	0.010	-0.026	-0.040
		(0.043)	(0.048)	(0.077)
(3)	Middle tercile	0.010	0.084	-0.058
		(0.033)	(0.036)	(0.063)
(4)	Top tercile	0.003	0.132	-0.062
		(0.045)	(0.053)	(0.068)
(5)	PD * Middle tercile	-0.062	0.033	0.052
		(0.042)	(0.044)	(0.090)
(6)	PD * Top tercile	-0.071	0.003	0.134
		(0.050)	(0.054)	(0.096)
(7)	PD + Follow up * Middle tercile	0.009	0.045	0.007
		(0.041)	(0.046)	(0.093)
(8)	PD + Follow up * Top tercile	-0.030	0.053	0.125
		(0.049)	(0.056)	(0.096)
(9)	Observations	14,599	14,599	14,599
Panel B: Comparing PD + Evaluation versus PD (left-out group)				
(10)	PD + Evaluation	0.037	0.018	0.063
		(0.044)	(0.050)	(0.067)
(11)	Middle tercile	-0.005	0.100	-0.020
		(0.034)	(0.040)	(0.065)
(12)	Top tercile	-0.071	0.144	0.115
		(0.052)	(0.059)	(0.067)
(13)	PD + Evaluation* Middle tercile	-0.037	0.031	0.006
		(0.039)	(0.047)	(0.087)
(14)	PD + Evaluation * Top tercile	0.019	0.005	-0.090
		(0.048)	(0.057)	(0.091)
(15)	Observations	9,726	9,726	9,726
Panel C: Comparing PD + Evaluation versus Control (left-out group)				
(16)	PD + Evaluation	0.034	-0.017	-0.012
		(0.040)	(0.044)	(0.071)
(17)	Middle tercile	0.000	0.063	-0.048
		(0.034)	(0.038)	(0.060)
(18)	Top tercile	-0.015	0.097	-0.061
		(0.048)	(0.059)	(0.066)
(19)	PD + Evaluation * Middle tercile	-0.042	0.056	0.008
		(0.040)	(0.043)	(0.087)
(20)	PD + Evaluation * Top tercile	-0.033	0.026	0.064
		(0.046)	(0.050)	(0.092)
(21)	Observations	10,006	10,006	10,006

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for student and teacher baseline covariates and block fixed effects. PD stands for professional development. According to the standard error estimates, none of the coefficients are statistically significant at the 10 percent level and this of course remains when adjusting p-values for multiple hypothesis testing.

TABLE 7 – IMPACTS ON STUDENT ACHIEVEMENT BY TEACHER CHARACTERISTICS (AT ENDLINE)

		Female (yes/no) (1)	College degree (yes/no) (2)	Math major (yes/no) (3)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)				
(1)	PD	0.020	0.055	-0.024
		(0.049)	(0.042)	(0.042)
(2)	PD + Follow-up	-0.004	0.097	0.049
		(0.049)	(0.041)	(0.041)
(3)	Group	0.071	0.122	0.022
		(0.051)	(0.052)	(0.052)
(4)	PD * Group	-0.051	-0.203	0.049
		(0.069)	(0.074)	(0.070)
(5)	PD + Follow-up * Group	0.020	-0.312	-0.143
		(0.070)	(0.078)	(0.072)
(6)	Observations	14,599	14,599	14,599
Panel B: Comparing PD + Evaluation versus PD (left-out group)				
(7)	PD + Evaluation	0.033	0.041	0.035
		(0.051)	(0.041)	(0.043)
(8)	Group	0.053	-0.170	-0.010
		(0.055)	(0.081)	(0.062)
(9)	PD + Evaluation * Group	-0.004	-0.035	-0.014
		(0.072)	(0.083)	(0.077)
(10)	Observations	9,726	9,726	9,726
Panel C: Comparing PD + Evaluation versus Control (left-out group)				
(11)	PD + Evaluation	0.020	0.087	0.032
		(0.046)	(0.039)	(0.041)
(12)	Group	0.064	0.122	0.020
		(0.050)	(0.055)	(0.051)
(13)	PD + Evaluation * Group	-0.019	-0.254	-0.052
		(0.065)	(0.071)	(0.064)
(14)	Observations	10,006	10,006	10,006

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for teacher baseline covariates and block fixed effects. PD stands for professional development. We find significant effects for specific teacher subgroups even after adjusting p-values for multiple hypothesis testing. In particular, we find that relative to the control group: (a) PD plus follow-up and PD (only) have negative effects on the achievement of students whose teachers went to four year college (-0.215 and -0.147 SDs respectively, both significant at the 5 percent level, adjusted p-values of 0.003 and 0.035); (b) PD plus follow-up has small, positive effects on the achievement of students whose teachers did not go to four year college (0.097 SDs, significant at the 5 percent level, adjusted p-value = 0.047); (c) PD plus evaluation has negative effects on the achievement of students whose teachers went to four year college (-0.167 SDs, significant at the 5 percent level, adjusted p-value = 0.033) and smaller, positive effects on the achievement of students whose teachers did not go to four year college (0.087 SDs, significant at the 5 percent level, adjusted p-value = 0.004).