April 2017

Unpacking Teacher Professional Development

Prashant Loyalka, Anna Popova, Guirong Li, Chengfang Liu, and Henry Shi

Abstract

Despite massive investments in teacher professional development (PD) programs in developing countries, there is little evidence on their effectiveness. We present the results of a large-scale, randomized evaluation of a high-profile PD program in China, in which teachers were randomized to receive PD; PD plus follow-up; PD plus evaluation of their command of the PD content; or no PD. Precise estimates indicate that PD and associated interventions failed to improve teacher and student outcomes. 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.

Keywords: Teacher Quality, Professional Development, Follow-up, Evaluation, Randomized Trial, Developing Countries

JEL Codes: I24, O15, J33, M52

Working Paper 314

April 2017

reap.fsi.stanford.edu

REAP Rural Education Action Program

Introduction

Student achievement levels and gains are alarmingly low in developing countries (Kingdon, 2007; Das & Zajonc, 2010; Freeman et al., 2010; Glewwe & Muralidharan, 2015; UNESCO, 2015). Researchers have attributed these low levels of learning to a number of factors such as poor nutrition and health (Soemantri, Pollitt, & Kim, 1985; Soemantri, 1989; Luo et al., 2012; Kleiman-Weiner et al., 2013), insufficient educational inputs at home (Glewwe & Kremer, 2006; Glewwe & Miguel, 2008), as well as a lack of incentives to teach (Muralidharan & Sundararaman, 2011; Loyalka et al., 2016) and learn (Kremer et al., 2009).

Raising teacher quality has been shown to be one of the most important ways that educators can improve the learning of poorly performing students. Teacher quality, in both developing and developed countries, has consistently been shown to be closely associated with improvements in student learning (Rockoff, 2004; Hanushek & Rivkin, 2010; Chetty et al., 2014; Bruns & Luque, 2015). For example, the difference between a high and low quality teacher amounts to a difference of 0.3 standard deviations (SDs) on standardized tests in secondary school in Chile (MINEDUC, 2009) and to a full year of student learning in the United States (Hanushek & Rivkin, 2010). Teacher quality has also been shown to significantly improve long-term outcomes such as college graduation rates and adult salaries (Chetty et al., 2014).

Unfortunately, researchers have found that a large proportion of teachers in developing countries are ill-prepared for teaching (Villegas-Reimers, 1998; Ball, 2000). Teachers lack the requisite knowledge and skills to improve student achievement (Berhman et al., 1997; Behrman et al., 2008; Bruns & Luque, 2015; Tandon and Fukao, 2016; Bold et al., 2017). Despite sometimes high levels of formal education among teachers in developing countries, many exhibit weak cognitive skills and ineffective classroom practice. For example, across three Latin American countries, fewer than 3 percent of teachers score in the range considered excellent on tests of content mastery, and in no country do teachers engage the entire class more than 25 percent of the time (Bruns & Luque, 2015). In six African countries, only 10 percent of teachers score above the minimum for general pedagogical knowledge, and only 12 percent of teachers can comment on the learning progression of their students (Bold et al., 2017). Finally, in Cambodia, teachers score only slightly above ninth grade students in mathematics and score very low on tests of pedagogical content knowledge (Tandon and Fukao, 2015).

Aware of the role that high 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 aim of PD programs is to help existing teachers gain subject-specific knowledge and skills (Dadds, 2001), use appropriate instructional practices (Darling-Hammond & McLaughlin, 1995; Schifter et al., 1999), develop positive attitudes and values, and ultimately improve student learning (Villegas-Reimers, 2003). Since subject-specific knowledge and skills (Hill et al., 2005; Metzler & Woessman, 2011; Shepard, 2015; Bold et al., 2017), appropriate instructional practices (Rowan et al., 2002; Hiebert & Grouws, 2007), and positive changes in values and attitudes (Stern & Shavelson, 1983; Fang, 1996) have strong positive associations with student achievement in developing countries, the policy to promote teacher PD appears to have a strong logical basis.

There are at least four reasons, however, why teacher PD programs may fail to improve teacher and student outcomes. First, the content of PD programs themselves may be of low quality and/or not relevant to the practical concerns of teachers (Castro, 1991; Subirats & Nogales, 1989). Second, while the content may be appropriate, the delivery of PD programs may be ineffective (Villegas-Reimers, 1998; Villegas-Reimers, 2003). Third, teachers that go through PD programs may fail to implement what they learned in the programs due to insufficient follow-up (Cohen, 1990; Lieberman, 1994; Corcoran, 1995; Guskey, 1995; Schifter, Russell, & Bastable 1999, p. 30; Dudzinski, 2000; Ganser 2000; Villegas-Reimers, 2003). In other words, teachers may learn knowledge and skills during an initial set of training sessions but require follow-up to reinforce this learning and translate it into practice. Fourth, even if teachers are able to acquire knowledge and skills from teacher PD programs, they may fail to hold trainees accountable for improving their teaching habits (Subirats & Nogales, 1989; Braslavsky & Birgin, 1992). In other words, teachers may require a combination of incentives, evaluation and feedback to ensure they put what they learned in PD programs into practice (Guskey, 1995). Taken together, these potential weaknesses in the design and implementation of teacher PD programs, may undermine impacts on teaching and learning. Since teacher PD programs further require teachers, school administrators and policymakers to substitute time and resources away from students, they may even lead to negative impacts.

The effectiveness of teacher PD is thus an empirical question. Evidence from high-income countries generally shows that teacher PD is effective at improving student achievement and

points towards PD that includes detailed instructions on implementation, follow-up support, and significant contact hours, as being more effective at raising student test scores (Yoon et al., 2007; Fryer, 2016). However, there is considerable variation in effect sizes across programs, with some program evaluations even showing negative effects. Moreover, there is substantial variation in the quality of studies from which these results are drawn.[1]

Evidence from developing countries is yet more limited. Despite the importance that is being placed on PD and the fact that billions of dollars and billions of teacher hours are being invested in PD programs each year, evidence on the effectiveness of the programs is lacking (OECD, 2009; Bruns & Luque, 2015).[2] In fact, the limitations of the empirical evidence on the effectiveness of PD programs are threefold. First, there have been almost no large-scale randomized evaluations of teacher PD programs on student achievement in developing countries.[3] Second, to the best of our knowledge, there are few large-scale randomized evaluations in developed or developing countries that examine whether specific design features of teacher PD programs such as post-training follow-up and evaluation are effective. Finally, few randomized evaluations from either developed or developing countries have systematically studied the causal pathway through which teacher PD programs impact, or fail to impact, teacher

[1] 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). 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, an experimental evaluation by the U.S. Department of Education comparing PD and PD plus coaching for early reading instruction found no significant impacts on student achievement (Garet et al., 2008). However, with only 30 schools in each treatment group, its power to test for design contrasts was limited. Their experimental evaluation of a middle school mathematics PD program also found no significant learning impacts after two years, but the finding was based on a sample of just 92 teachers (Garet et al., 2011).

[2] For example, between 2007 and 2016, Indonesia's national government allocated approximately 2.7 billion dollars (USD) to its Teacher Upgrading and Certification program (Jalal et al., 2009). Between 2012 and 2017, India's national government allocated approximately 1.2 billion USD to teacher PD programs (Government of India, 2011). In addition to these direct costs at the national level, local governments also spend considerable funds on teacher PD. Moreover, 63% of World Bank Education projects between 2000 and 2012 included professional development to support teachers (Popova et al., 2016). Finally, the indirect, opportunity costs of teacher PD are staggering. For example, teachers in Mexico spent an average of 34 days in teacher PD over 18 months (OECD, 2009).

[3] 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.

and student outcomes.[4] The absence of rigorous evidence along these dimensions 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.

Given these knowledge gaps, the overall purpose of this paper is to evaluate the impact of teacher PD on a wide range of teacher and student outcomes in a developing country context. We not only aim to examine the effectiveness of teacher PD but also the effectiveness of additional interventions such as post-training follow-up and evaluation that may increase the impact of PD. 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 in developing countries is to create a core group of teachers that can influence the teaching practices of other teachers (Gu, 1990; Darling-Hammond, Bullmaster, & Cobb, 1995; Cochran-Smith & Lytle, 1999; Berry, 2011; Zepeda, 2011), we also examine the degree to which PD programs have positive spillovers on peer teachers and students.[5]

To fulfill these goals, we conducted a large randomized evaluation of China's flagship national teacher PD program (*guojiaji peixun jihua* or *guopei* for short) and two accompanying post-training interventions that are believed to strengthen the impact of teacher PD. The post-training interventions consisted of: (a) continuous *follow-up* with trainees, alerting them of online supplementary materials, assignments, and progress reports through text messages and phone calls; and (b) an *evaluation* of how much trainees recalled from the PD program. Altogether we collected 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.

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 year. 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

[4] Three unpublished randomized evaluations conducted by the Institute of Education Sciences (IES) have studied the causal mechanisms driving impacts of teacher PD on student learning in reading and math across various grades in the U.S. (Garet et al., 2008, 2010, 2016).

[5] 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.

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, or on time spent on math. Nor does any combination of teacher PD with or without post-training follow-up or evaluation have any significant impact 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 identify two major reasons for 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 (as defined by not having a degree in the subject they are teaching and not having a college degree). 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, even low-quality 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.

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. When the content and delivery of PD is overly theoretical, adding design features such as follow-up or evaluation does little to improve its effectiveness. At best, heterogeneous responses to treatment from different teachers suggest that teacher PD programs may need to move beyond one-size-fits-all approaches. Our sample is large enough and sufficiently powered to identify even small effects, meaning the null findings should be taken seriously.

The rest of the paper proceeds as follows. Section II presents experimental design and data. Section III discusses the results, and Section IV concludes.

Experimental Design & Data

A. Sample

The study was conducted in Henan province in central China, in collaboration with the Provincial Department of Education. Henan is a lower income province, ranking 24 out of 31 provinces in terms of income per capita (NBS, 2015). It has a large population size of 94 million persons—if it were a country, it would rank it as the fourteenth largest in the world (NBS, 2011).

The Henan Provincial Department of Education provided a representative list of 300 rural junior high schools from 94 (out of 159) counties across the province and one grade 7-9 math teacher from each school to participate in the study.[6] 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.[7] 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.

B. 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

[6] 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 (Loyalka et al., 2015; Shi et al., 2015).

[7] 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.

"teacher PD plus follow-up" (treatment arm C in Figure 1).[8,9] 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.[10]

C. Intervention

China's government has invested heavily in teacher PD programs.[11] 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).[12] 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.

---

[8] 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).

[9] 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.

[10] Power calculations conducted using Optimal Design (Spybrook et al., 2009) indicate that with a pre-test intraclass correlation coefficient of =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.

[11] 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., 2015).

[12] 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.

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 training at a centralized location in November 2015. 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 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. Training providers conducted the follow-up through mobile text messages and phone calls. 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 training provider called the trainee to confirm he or she had

received the message.

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.

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 evaluation 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.[13]

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.

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.

[13] 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 & Park, 2017; Loyalka et al., 2015).

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)[14], and how 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 & Peterson, 1986; Fang, 1996; Kagan, 1992; Stipek et al., 2001; Thompson, 1992). [15]

Student Standardized Mathematics Tests. 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, tailored to the national and provincial-level mathematics curricula. Although grade-appropriate tests may present a problem in some developing countries (since

[14] 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). 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:

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

where 1 is a column vector of 1s, $\overline{\sum}_j^{-1}$ 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.

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

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.

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 et al., 2005), at baseline, midline, and endline. These were similarly normalized. Estimated effects are thus also expressed in standard deviations.

E. Balance and attrition

Appendix Tables A1 and A2 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. 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.[16]

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.[17] More importantly, cross-treatment differences in baseline student characteristics among non-attriters (Rows 1-2, 7, and 10 in Table A3) are virtually identical to cross-treatment

[16] 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).

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

differences in baseline student characteristics among the full baseline sample (Rows 1-2, 5, and 7 respectively in Table A2).[18] We therefore find no evidence of differential attrition across any of our treatment comparisons.

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.[19]

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

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 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.

[18] 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).

[19] 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.

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.

We also 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 (Table 1). 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 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 8). 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 8). 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 12). In fact, the upper limits of

the 95% 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.

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 8), 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 10), indicating that Evaluation has no additional effect beyond PD. The small point estimates in each of these cases lie within tight 95% confidence intervals, once again ruling out sizeable positive impacts.

We also find that PD and post-training components have no impacts on a wide range of secondary student outcomes (Table 2). 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 2, 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 2, 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 2, Panel A, Rows 3-4, Columns 1-9) or of Evaluation beyond PD (Table 2, 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. However, once we adjust for multiple hypothesis testing, the significance of these results falls away, with all adjusted p-values greater than 0.1.

The lack of positive effects on student outcomes is mirrored by the lack of impacts on (student-reported) teaching behaviors in the classroom (Table 3). 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).[20] Similarly, none of the individual PD components have significant effects on any measures of teacher behavior.

Having found no positive effects of PD and post-training components on teacher behaviors,

[20] 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.

we next examine whether they have any impact on teacher knowledge, attitudes, and beliefs. These may be important channels through which PD ultimately effects 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 & 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. On the one hand, individual results from Table 4 suggest that PD with Evaluation increases teacher math knowledge (Rows 2 and 8, Column 1), that PD and PD plus Evaluation decrease teachers' beliefs that student learning should be more heavily directed by the teacher (Rows 1 and 8, Column 4), and that the Evaluation component of PD (over PD alone) increases teachers' intrinsic motivation (Row 6, Column 2). On the other hand, PD alone decreases teachers' beliefs that math should be taught more actively (Row 1, Column 5), and PD plus Follow-up increases erroneous beliefs that math is a fixed (and not learned) ability (Row 2, Column 6). Furthermore, none of the 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 5), 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.

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).[21] We do, however, find some variation in effects by teacher qualifications (Table 7). Namely, PD significantly decreases scores among students whose teachers had a college degree relative to

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

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—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).

Results are similar even after we adjust standard error estimates for multiple hypothesis testing.[22] 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% level); (b) PD plus follow-up has small, positive effects on the achievement of students whose teachers did not go to four year college (.097 SDs, significant at the 5% level); (c) PD plus evaluation has negative effects on the achievement of students whose teachers went to four year college (-.167 SDs, significant at the 5% level) 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% level).

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 it causes all teachers to substitute time away from teaching. However, if more qualified teachers were originally helping students learn, while less qualified teachers were perhaps not contributing to learning, then only a loss in the teaching time of qualified teachers would have negative consequences for student learning.

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 (including fundamental attitudes about the nature of math and math teaching), and behaviors (including teaching practices) 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

[22] Although we stated in our pre-analysis plan that we would not adjust the standard errors in the heterogeneous effects analysis for multiple hypothesis testing (since we treat the analyses as exploratory), we nonetheless adjusted the standard error estimates for the fact that we tested 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.

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 indeed 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.[23]

Trainee participation was high. According to our records, daily attendance for the on-site PD sessions was 93 percent.[24] 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.[25] 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 47 percent of the materials were "extremely theoretical" with little application to the real world. 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

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

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

[25] 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.

"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.

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. A large number 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 practical applications they did learn from PD in their classrooms. Several 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. Some teachers also noted that the large degree of heterogeneity in student ability in their classrooms also prohibited them from applying new teaching techniques.

In summary, the accumulated evidence suggests that the biggest reason for the failure of the teacher PD program lies in its content and delivery. Not only did teachers describe the content as overly theoretical and the delivery as rote, but they also clearly did not learn new math knowledge or change their beliefs about the nature of math teaching (either by the time of the midline or the endline survey). This was despite the fact that an explicit major goal of the teacher PD program was to increase teacher math knowledge and their understanding about how to teach math effectively. Given that PD and its associated post-training components failed to affect teacher math knowledge and beliefs about math teaching, it is little wonder then that they had few effects on the more distal parts of the causal chain, such as teaching practices and student

outcomes.

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 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. Based on data from teacher interviews, we attribute this lack of impact to the fact that PD content was overly theoretical and its delivery was rote and passive, making it difficult for teachers to remember, relate to, and implement. 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, pinpointing reasons for the lack of impacts. Fourth and finally, this is the first large-scale experimental evaluation of a government-sponsored teacher PD policy in a developing country. Most experimental evaluations of teacher PD programs in developing countries are efficacy studies that are implemented with a high degree of fidelity (Gartlehner et al., 2006), 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. 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, potentially offer additional insights for policymakers. Lower cost solutions–such as online PD sessions and follow-up, follow-up reminders via text message or phone call, and one-off evaluations may do little to increase the effectiveness of PD. Instead, educational researchers would argue that more human resource intensive follow-up (mentoring visits) and evaluation (formative assessment) would be more effective (Popova, Evans & Arancibia, 2016; Hobson et al., 2009; Guskey, 2002). 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 in China alone, policymakers may wish to consider investing in other types of PD programs that find more support in education

theory and practice.[26] 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.

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

REFERENCES

Abeberese, A. B., Kumler, T. J., & Linden, L. (2012). *Improving reading skills by encouraging children to read: A randomized evaluation of the Sa Aklat Siskat reading program in the Philippines*. Unpublished manuscript, Columbia University, New York, NY

Alderman, H., Behrman, J. R., Khan, S., Ross, D. R., & Sabot, R. (1996). Decomposing the Regional Gap in Cognitive Skills in Rural Pakistan. *Journal of Asian Economics*, *7*(1), 49-76.

Anderson, M.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, 1481–1495.

Ball, A. F. (2000). Preparing teachers for diversity: Lessons learned from the US and South Africa. *Teaching and Teacher Education*, *16*(4), 491-509.

Behrman, J. R., Khan, S., Ross, D., & Sabot, R. (1997). School Quality and Cognitive Achievement Production: A Case Study for Rural Pakistan. *Economics of Education Review*, *16*(2), 127-142.

Behrman, J. R., Ross, D., & Sabot, R. (2008). Improving Quality versus Increasing the Quantity of Schooling: Estimates of Rates of Return from Rural Pakistan. *Journal of Development Economics*, *85*(1), 94-104.

Berry, B. (Ed.). (2011). *Teaching 2030: What we must do for our students and our public schools: Now and in the future.* Teachers College Press.

Bold, T., Filmer, D., Martin, G., Molina, E., Rockmore, C., Stacy, B., Svensson, J., Wane, W. (2017). What Do Teachers Know and Do? Does It Matter? World Bank Policy Research Working Paper 7956.

Bruhn, M., McKenzie, D., 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." A*merican Economic Journal: Applied Economics 1*(4): 200–232.

Bruns, B., & Luque, J. (2015). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: The World Bank.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: teacher and health worker absence in developing countries. *The Journal of Economic Perspectives,* 20(1), 91-116.

Chetty, R, JN Friedman, JE Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review, 104*(9): 2633-79.

Clark, C.M., & Peterson, P. L. (1986). Teachers' thought processes. In M. Wittrock (Ed.), Handbook of research in teaching (3rd ed.), MacMillan, New York (1986), pp. 255–296.

Cochran-Smith, M., & Lytle, S. L. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education, 24*, 249-305.

Darling-Hammond, L., Bullmaster, M. L., & Cobb, V. L. (1995). Rethinking teacher leadership through professional development schools. *The Elementary School Journal, 96*(1), 87-106. doi: 0013-5984/96/9601-0006

Das, J., & Zajonc, T. (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.

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, *38*(1), 47-65.

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

Freeman, R. B., Machin, S., & Viarengo, M. (2010). Variation in Educational Outcomes and Policies across Countries and of Schools within Countries. National Bureau of Economic Research W16293.

Fryer Jr, R. G. (2016). *The production of human capital in developed countries: Evidence from 196 randomized field experiments* (NBER Working Paper No. 22130). Cambridge, MA: National Bureau of Economic Research.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pdf/20084030.pdf.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., et al. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher

professional development (NCEE 2016-4010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20164010/pdf/20164010.pdf

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P. & Sepanik, S. (2011). Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation. NCEE 2011-4024. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., et al. (2010). Middle school mathematics professional development impact study: Findings after the first year of implementation (NCEE 2010-4009). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20104009/pdf/20104009.pdf

Gersten, R., Taylor, M. J., Keys, T.D., Rolfhus, E., and Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches.* Washington, DC: 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.

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 945–1017). Amsterdam, Netherlands: Elsevier.

Glewwe, P., & Miguel, E. A. (2008). The impact of child health and nutrition on education in less developed countries. In T. P. Schultz & J. Strauss (Eds.), *Handbook of development economics* (*Vol. 4*, pp. 3561 – 3606)

Glewwe, P. & Muralidharan, K. (2015). *Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications.* Handbook of the Economics of Education. Volume 5.

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.

Gu, M. (1990). *Zhongguo jiaoyu cidian* [*Chinese Educational Dictionary*]. Shanghai, China:

Shanghai Education Press.

Guskey, T.R. (2002). Professional Development and Teacher Change, *Teachers and Teaching*, 8:3, 381-391.

Hanushek, E. & Rivkin, S. 2010. Using Value-Added Measures of Teacher Quality. CALDER.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. *Second handbook of research on mathematics teaching and learning*, 1, 371-404.

Hobson, A. J., Ashby, P., Malderez, A., & Tomlinson, P. D. (2009). Mentoring beginning teachers: What we know and what we don't. *Teaching and Teacher Education*, 25(1), 207-216.

Jacob, R., Hill, H., & Corey, D. (2017). The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379-407.

Jalal, F., Samani, M., Chang, M. C., Stevenson, R., Ragatz, A. B., & Negara, S. D. (2009). Teacher certification in Indonesia: A strategy for teacher quality improvement. Jakarta: Ministry of National Education Indonesia/The World Bank.

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist*, *27*(1), 65-90.

Karachiwalla, N. and Park, A. (2017). Promotion Incentives in the Public Sector: Evidence from Chinese Schools. *Journal of Public Economics*, Forthcoming.

Kingdon, G. G. (2007). The Progress of School Education in India. *Oxford Review of Economic Policy*, 23(2), 168-195.

Kleiman-Weiner, M., Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, S. (2013). Eggs versus chewable vitamins: Which intervention can increase nutrition and test scores in rural China? *China Economic Review*, *24*, 165–176.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *Review of Economics and Statistics*, *91*, 437–456.

Laschke, C., & Blömeke, S. (Eds.). (2014). *Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M 2008). Dokumentation der Erhebungsinstrumente*. Waxmann Verlag.

Lee, D.S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on

Treatment Effects. *Review of Economic Studies,* 76, 1071–1102.

Li, Y. (2013) "The Situation and Reflection on the Implementation of Inner Mongolia National Training Plan" *Journal of Inner Mongolia Normal University* Vol. 12 No. 26, pp. 63-97. (in Mandarin Chinese).

Loyalka, P., Sylvia, S., Liu, C.F., Chu, J., Shi, Y.J. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." REAP Working Paper. Forthcoming.

Luo, R., Shi, Y., Zhang, L., Liu, C., Rozelle, S., Sharbono, B….Martorell, R. (2012). Nutrition and educational performance in rural China's elementary schools: Results of a randomized control trial in Shaanxi Province. *Economic Development and Cultural Change*, *60*, 735 – 772.

Ministry of Education (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.

MINEDUC (Ministerio de Educación, Chile). 2009. Resultados Nacionales SIMCE 2008. Santiago: MINEDUC.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*, 39-77.

Murnane, R. J., & Ganimian, A. J. (2014). *Improving educational outcomes in developing countries: Lessons from rigorous evaluations* (No. w20284). National Bureau of Economic Research.

National Bureau of Statistics. 2010. China Statistical Yearbook. Beijing: China National Bureau of Statistics.

National Bureau of Statistics. 2011. "Communiqué of the National Bureau of Statistics of People's Republic of China on Major Figures of the 2010 Population Census [1] (No. 2)". Retrieved 4 August 2013.

National Bureau of Statistics. 2015. China Statistical Yearbook. Beijing: China National Bureau of Statistics.

Nordstrom., L. E. (2013). Teacher supply, training and cost in the context of rapidly expanding enrolment: Ethiopia, Pakistan and Tanzania. Background paper prepared for the

Education for All Global Monitoring Report 2013/4/ Teaching and learning: Achieving quality for all. United Nations Educational, Scientific and Cultural Organization.

Nitsaisook, M., & Anderson, L. W. (1989). An experimental investigation of the effectiveness of inservice teacher education in Thailand. *Teaching and Teacher Education*, *5*, 287-302.

OECD. (2009). Creating Effective Teaching and Learning Environments: First Results from TALIS. Teaching and Learning International Survey. Organisation for Economic Co-Operation and Development.

Popova, A., Evans, D. K., & Arancibia, V. (2016). *Training Teachers on the Job: What Works and How to Measure it.* (World Bank Policy Research Working Paper No. 7834). Washington, DC: World Bank.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 417-458.

Rockoff, J. E. (2004). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review*, 94(2): 247–52.

Rowan, B., Correnti, R., & Miller, R. (2002). What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools. *The Teachers College Record*, 104(8), 1525-1567.

Sargent, T. C. (2014). Professional learning communities and the diffusion of pedagogical innovation in the Chinese education system. Comparative Education Review, 59(1), 102-132.

Shepherd, D. (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 (No. 01/2015).

Shi, Y., L. Zhang, Y. Ma, H. Yi, C. Liu, N. Johnson, J. Chu, P. Loyalka and S. Rozelle (2015). Dropping out of rural China's secondary schools: A mixed-methods analysis." *The China Quarterly* 224: 1048-1069.

Soemantri, A. G. (1989). Preliminary findings on iron supplementation and learning achievement of rural Indonesian children. *American Journal of Clinical Nutrition*, *50*, 698–702.

Soemantri, A. G., Pollitt, E., & Kim, I. (1985). Iron deficiency anemia and educational achievement. *American Journal of Clinical Nutrition*, *42*, 1221–1228.

Spybrook, J., Raudenbush, S.W., Liu, X., Congden, R. & Martinez, A. (2009). Optimal Design

for Longitudinal and Multilevel Research v1.76 [Computer Software].

Stern, P. and Shavelson, R. (1983). Reading teachers' judgements, plans, and decision making. *Reading Teacher*, 37, 280-6.Vegas, E. (2007). Teacher Labor Markets in Developing Countries. *The Future of Children*, 17(1), 219-232.

Stipek, D. J., Givvin, K. B., Salmon, J. M., & MacGyvers, V. L. (2001). Teachers' beliefs and practices related to mathematics instruction. *Teaching and teacher education*, *17*(2), 213-226.

Tandon, P., & Fukao, T. (2015). Educating the next generation: Improving teacher quality in Cambodia. World Bank Publications.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D.A Grouws (Ed.), Handbook of research on mathematics teaching and learning, Macmillan, New York (1992), pp. 127–146.

UNESCO. (2015). Fixing the broken promise of education for all: Findings from the Global Initiative on Out-of-School Children. UNESCO Institute for Statistics and United Nations Children's Fund.

Villegas-Reimers, E. (1998). The Preparation of Teachers in Latin America: Challenges and Trends. Human Development Department, World Bank, Latin America and the Caribbean Regional Office.

Villegas-Reimers, E. (2003). *Teacher PD: an international review of the literature*. Paris: International Institute for Educational Planning.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, *54*(3), 427-450.

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

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher PD Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

Zepeda, S. J. (2011). *Professional development: What works* (2nd ed.). Larchmont, NY: Eye on Education.

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

602.

Zuo E. and Su, Q. (2012), “The Investigation of Rural Middle School Core Physical Education Teachers Participating in Hunan Province ‘National Training Program’.” *Nenjiang Technology,* Vol. 3 No. 2, pp. 20-25.

IV. Figures

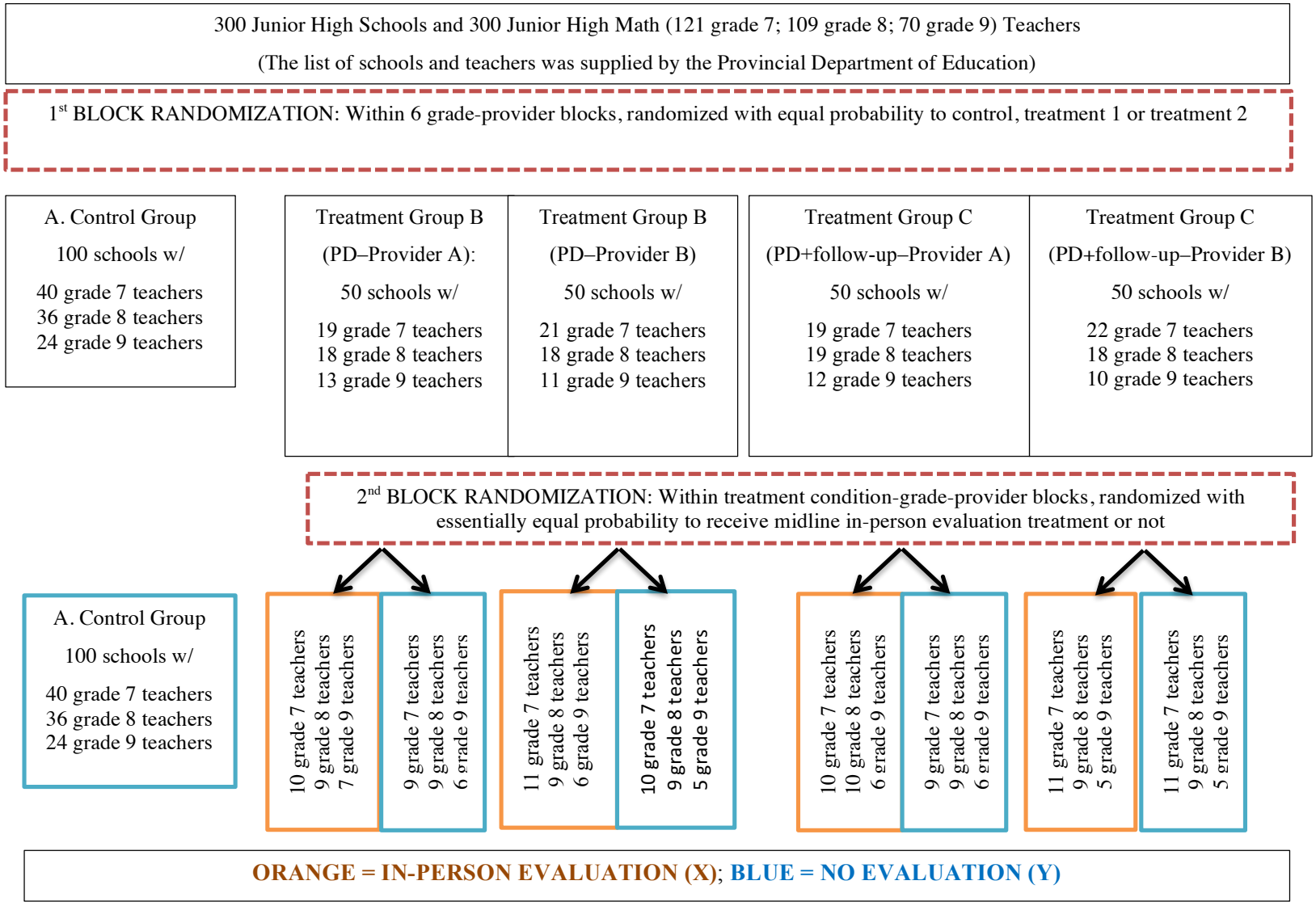

FIGURE 1. RANDOMIZATION PROCEDURE.

V. Tables

TABLE 1 – IMPACTS ON STUDENT ACHIEVEMENT (AT MIDLINE AND ENDLINE)

		Midline						Endline					
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)													
(1)	PD	-0.015	-0.035					0.023	-0.006				
		(0.028)	(0.027)					(0.036)	(0.034)				
(2)	PD + Follow-up	0.000	-0.020					0.026	0.005				
		(0.031)	(0.030)					(0.037)	(0.035)				
(3)	Difference: PD + Follow-up - PD	0.015	0.015					0.003	0.012				
(4)	P-value: PD + Follow-up - PD	0.609	0.613					0.934	0.749				
(5)	Observations	15,987	15,713					14,838	14,599				
Panel B: Comparing PD + Evaluation versus PD (left-out group)													
(6)	PD + Evaluation			0.008	0.005					0.043	0.031		
				(0.029)	(0.028)					(0.037)	(0.034)		
(7)	Observations			10,725	10,483					9,934	9,726		
Panel C: Comparing PD + Evaluation versus Control (left-out group)													
(8)	PD + Evaluation					-0.003	-0.022					0.044	0.011
						(0.028)	(0.028)					(0.035)	(0.032)
(9)	Observations					10,967	10,774					10,168	10,006
(10)	Additional controls		X		X		X		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. *** p<0.01, ** p<0.05, * p<0.1

TABLE 2 - 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)
(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
(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
(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. *** p<0.01, ** p<0.05, * p<0.1. After adjusting p-values for multiple hypothesis testing using the Free Step-Down Resampling Method (Westfall & Young, 1993), none of the estimated coefficients are significant at the 10 percent level.

TABLE 3 - 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. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 4 - 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 that math is, by nature, 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. *** p<0.01, ** p<0.05, * p<0.1. After adjusting p-values for multiple hypothesis testing using the Free Step-Down Resampling Method (Westfall & Young, 1993), none of the estimated coefficients are significant at the 10 percent level.

TABLE 5 - IMPACTS ON STUDENT ACHIEVEMENT IN SPILLOVER SAMPLE (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	-0.033	-0.025				
		(0.040)	(0.034)				
(2)	PD + Follow-up	-0.018	-0.010				
		(0.041)	(0.036)				
(3)	Difference: PD + Follow-up - PD	0.014	0.014				
(4)	P-value: PD + Follow-up - PD	0.710	0.707				
(5)	Observations	15,173	14,789				
Panel B: Comparing PD + Evaluation versus PD (left-out group)							
(6)	PD + Evaluation			0.070*	0.057		
				(0.038)	(0.037)		
(7)	Observations			10,332	10,050		
Panel C: Comparing PD Evaluation versus Control (left-out group)							
(8)	PD + Evaluation					0.007	0.018
						(0.039)	(0.033)
(9)	Observations					10,288	10,093
(10)	Additional controls		X		X		X

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for student and teacher baseline covariates and block fixed effects. PD stands for professional development.
*** p<0.01, ** p<0.05, * p<0.1

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

		Student household wealth (asset index) (1)	Student baseline achievement (2)	Hours teacher received PD previous to the baseline (3)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)				
(1)	PD	0.035	-0.021	-0.067
		(0.045)	(0.054)	(0.073)
(2)	PD + Follow-up	0.010	0.003	-0.040
		(0.043)	(0.056)	(0.077)
(3)	Middle tercile	0.010	0.870***	-0.058
		(0.033)	(0.035)	(0.063)
(4)	Top tercile	0.003	1.493***	-0.062
		(0.045)	(0.043)	(0.068)
(5)	PD * Middle tercile	-0.062	0.030	0.052
		(0.042)	(0.051)	(0.090)
(6)	PD * Top tercile	-0.071	-0.004	0.134
		(0.050)	(0.062)	(0.096)
(7)	PD + Follow up * Middle tercile	0.009	0.011	0.007
		(0.041)	(0.054)	(0.093)
(8)	PD + Follow up * Top tercile	-0.030	0.012	0.125
		(0.049)	(0.064)	(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	0.016	0.063
		(0.044)	(0.058)	(0.067)
(11)	Middle tercile	-0.005	0.869***	-0.020
		(0.034)	(0.039)	(0.065)
(12)	Top tercile	-0.071	1.483***	0.115*
		(0.052)	(0.051)	(0.067)
(13)	PD + Evaluation* Middle tercile	-0.037	0.031	0.006
		(0.039)	(0.055)	(0.087)
(14)	PD + Evaluation * Top tercile	0.019	0.004	-0.090
		(0.048)	(0.064)	(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	-0.005	0.016
		(0.040)	(0.054)	(0.058)
(17)	Middle tercile	0.000	0.868***	0.869***
		(0.034)	(0.035)	(0.039)
(18)	Top tercile	-0.015	1.491***	1.483***
		(0.048)	(0.043)	(0.051)
(19)	PD + Evaluation * Middle tercile	-0.042	0.037	0.031
		(0.040)	(0.052)	(0.055)
(20)	PD + Evaluation * Top tercile	-0.033	0.002	0.004
		(0.046)	(0.061)	(0.064)
(21)	Observations	10,006	10,006	9,726

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for student and teacher baseline covariates and block fixed effects. PD stands for professional development. *** p<0.01, ** p<0.05, * p<0.1. As specified in our pre-analysis plan, when estimating heterogeneous effects, we do not adjust 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. *** p<0.01, ** p<0.05, * p<0.1. Significant effects remain significant at least at the 5% level even after adjusting p-values for multiple hypothesis testing (results not shown but available upon request).

APPENDIX A: SUPPLEMENTARY TABLES

TABLE A1 – BALANCE TESTS USING BASELINE TEACHER CHARACTERISTICS

		Teacher Female (yes/no)	Teacher Age (years)	University degree (yes/no)	Teacher has higher rank (yes/no)	Teacher has teaching certificate (yes/no)	Teacher experience (years)	Teacher majored in math (yes/no)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)								
(1)	PD	-0.030	-0.088	0.006	0.094	0.010	0.257	-0.017
		(0.071)	(1.047)	(0.066)	(0.070)	(0.010)	(1.231)	(0.069)
(2)	PD + Follow-up	-0.050	-1.066	-0.036	0.106	-0.000	1.407	-0.092
		(0.071)	(1.001)	(0.065)	(0.069)	(0.014)	(1.188)	(0.068)
(3)	Difference: PD + Follow-up - PD	-0.020	-0.978	-0.042	0.012	-0.010	1.150	-0.076
(4)	P-value: PD + Follow-up - PD	0.781	0.329	0.516	0.859	0.324	0.341	0.259
(5)	Observations	298	298	298	298	298	298	298
Panel B: Comparing PD + Evaluation versus PD (left-out group)								
(5)	PD + Evaluation	-0.042	-0.164	0.042	0.055	0.011	-0.166	0.111*
		(0.072)	(1.016)	(0.066)	(0.071)	(0.011)	(1.229)	(0.067)
(6)	Observations	198	198	198	198	198	198	198
Panel C: Comparing PD + Evaluation versus Control (left-out group)								
(7)	PD + Evaluation	-0.059	-0.752	0.001	0.131*	0.010	0.867	0.004
		(0.070)	(1.006)	(0.066)	(0.068)	(0.010)	(1.180)	(0.070)
(8)	Observations	202	202	202	202	202	202	202

Notes: Cluster-robust SEs in parentheses. Estimates are adjusted for block fixed effects. PD stands for professional development. *** p<0.01, ** p<0.05, * p<0.1

TABLE A2 – BALANCE TESTS USING BASELINE STUDENT CHARACTERISTICS

		Baseline Achievement (SDs)	Age (years)	Female (yes/no)	Father completed junior high or above (yes/no)	Mother completed junior high or above (yes/no)	Household wealth index (SDs)
		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)							
(1)	PD	-0.016	0.084**	-0.002	0.025	0.014	-0.004
		(0.068)	(0.042)	(0.011)	(0.022)	(0.027)	(0.075)
(2)	PD + Follow-up	-0.004	0.032	-0.008	-0.011	-0.017	0.063
		(0.072)	(0.048)	(0.011)	(0.024)	(0.029)	(0.071)
(3)	Difference: PD + Follow-up - PD	0.012	-0.052	-0.006	-0.036	-0.032	0.066
(4)	P-value: PD + Follow-up - PD	0.862	0.234	0.574	0.113	0.261	0.366
(5)	Observations	16,632	16,613	16,640	16,657	16,654	16,579
Panel B: Comparing PD + Evaluation versus PD (left-out group)							
(5)	PD + Evaluation	0.075	0.067	-0.014	0.008	0.015	-0.065
		(0.066)	(0.043)	(0.010)	(0.022)	(0.027)	(0.071)
(6)	Observations	11,153	11,136	11,164	11,176	11,173	11,125
Panel C: Comparing PD + Evaluation versus Control (left-out group)							
(7)	PD + Evaluation	0.025	0.091**	-0.012	0.012	0.007	-0.001
		(0.069)	(0.043)	(0.011)	(0.022)	(0.027)	(0.073)
(8)	Observations	11,401	11,392	11,389	11,402	11,401	11,345

Notes: The above balance tests are for "non-spillover classes". The balance test results are substantively the same when spillover classes are also included (i.e. there does not appear to be significant imbalance across any of the covariates). Cluster-robust SEs in parentheses. Estimates are adjusted for block fixed effects. PD stands for professional development. *** p<0.01, ** p<0.05, * p<0.1

TABLE A3 – ATTRITION BY BASELINE STUDENT CHARACTERISTICS

	Baseline Achievement (SDs)	Age (years)	Female (yes/no)	Father completed junior high or above (yes/no)	Mother completed junior high or above (yes/no)	Household wealth index (SDs)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)						
(1) PD	-0.021	0.081**	-0.004	0.021	0.009	-0.015
	(0.067)	(0.040)	(0.011)	(0.022)	(0.027)	(0.076)
(2) PD + Follow-up	-0.001	0.045	-0.009	-0.014	-0.016	0.056
	(0.070)	(0.046)	(0.012)	(0.024)	(0.030)	(0.073)
(3) Endline attrition	-0.661***	-0.366***	-0.147***	-0.085***	-0.060**	-0.031
	(0.075)	(0.054)	(0.028)	(0.026)	(0.026)	(0.068)
(4) PD # Endline attrition	0.047	0.042	0.019	0.074*	0.057	0.136
	(0.113)	(0.086)	(0.036)	(0.038)	(0.041)	(0.094)
(5) PD + Follow-up # Endline attrition	-0.038	-0.129*	0.028	0.028	-0.055	0.102
	(0.105)	(0.071)	(0.038)	(0.039)	(0.037)	(0.097)
(6) Observations	16,632	16,613	16,640	16,657	16,654	16,579
Panel B: Comparing PD + Evaluation versus PD (left-out group)						
(7) PD + Evaluation	0.082	0.066	-0.011	0.012	0.020	-0.082
	(0.064)	(0.041)	(0.011)	(0.023)	(0.028)	(0.073)
(8) Endline attrition	-0.660***	-0.442***	-0.110***	-0.037	-0.050	0.048
	(0.069)	(0.051)	(0.028)	(0.032)	(0.031)	(0.060)
(9) PD + Evaluation # Endline attrition	-0.028	0.042	-0.024	-0.001	-0.026	0.082
	(0.111)	(0.080)	(0.036)	(0.043)	(0.044)	(0.098)
Observations	11,153	11,136	11,164	11,176	11,173	11,125
Panel C: Comparing PD + Evaluation versus Control (left-out group)						
(10) PD + Evaluation	0.026	0.095**	-0.012	0.010	0.008	-0.018
	(0.068)	(0.042)	(0.011)	(0.022)	(0.028)	(0.074)
(11) Endline attrition	-0.669***	-0.362***	-0.148***	-0.084***	-0.057**	-0.029
	(0.075)	(0.054)	(0.028)	(0.026)	(0.026)	(0.068)
(12) PD + Evaluation # Endline attrition	0.008	-0.029	0.016	0.045	-0.017	0.158
	(0.114)	(0.083)	(0.036)	(0.037)	(0.039)	(0.100)
Observations	11,401	11,392	11,389	11,402	11,401	11,345

Notes: The above attrition tests are for "non-spillover classes". Cluster-robust SEs in parentheses. Estimates are adjusted for block fixed effects. PD stands for professional development. *** p<0.01, ** p<0.05, * p<0.1