

# 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

## Online Appendix

### Online Appendix A: China's National Teacher Training Program (NTTP): Background, Content, and Comparison with At-Scale Training Programs from Other Developing Countries

#### A. Content of the NTTP

Policymakers in China have mandated that the NTTP should cover three standard content areas: teacher knowledge, instructional skills (pedagogy), and professional values. Specifically, policy guidelines required that approximately 40% of NTTP content focus on teacher knowledge, 50% focus on instructional skills, and 10% focus on professional values. Table A1, immediately below, provides more information about the percentage of time devoted to the three content areas, as well as sub-content areas, within the NTTP that we observed and evaluated in this study:

TABLE A1. PERCENTAGE OF NTTP TIME DEVOTED TO DIFFERENT CONTENT AREAS

Area	Percentage of Time
<i>Knowledge</i>	
Subject matter knowledge	24%
Child and youth development	8%
Working with modern educational technologies	8%
<i>Pedagogy</i>	
Designing instruction	21%
Implementing instruction	11%
Diagnosing instruction and student performance	5%
Observing model classrooms	11%
<i>Values</i>	
Teacher ethics	8%
Teacher professionalism	5%

We also analyzed the content of the NTTP according to its focus on theoretical, practical or

miscellaneous material. This involved several steps. First, we obtained the detailed training modules of the 15-day onsite training program. We then coded distinct sections of the modules as being either theoretical, practical, or miscellaneous. A section was regarded as practical if it was largely characterized by content that the trainee could directly use for class preparation, class management, instruction, or student mentoring. A section was regarded as theoretical if it contained content that could not be applied, in any direct way, to pedagogical activities. A section was regarded as miscellaneous if its purpose was to organize the remaining material. Finally, the ratings of two coders were checked for consistency and resolved by consultation in the case of differences. The resulting percentages of theoretical, practical, and miscellaneous content, by module and in aggregate, are shown in Table A2 immediately below.

TABLE A2. PERCENTAGE OF NTTP CONTENT DEVOTED TO THEORETICAL AND PRACTICAL MATERIAL

Module No.	Percentage of theoretical material (sections)	Percentage of practical material (sections)	Miscellaneous material (organizational, etc.)
1	82.8%	14.5%	2.6%
2	20.2%	77.0%	2.8%
3	30.5%	67.5%	2.0%
4	71.0%	22.1%	6.9%
5	49.0%	47.8%	3.2%
6	56.0%	36.0%	8.0%
7	46.2%	57.2%	2.1%
8	27.7%	69.7%	2.6%
9	19.1%	75.5%	5.3%
10	25.5%	68.6%	5.9%
11	93.1%	3.4%	3.4%
12	97.0%	0.0%	3.0%
13	93.3%	0.0%	6.7%
14	98.6%	0.0%	1.4%
15	71.1%	24.4%	4.4%
16	62.2%	35.7%	2.1%
17	5.0%	87.5%	7.5%
18	8.9%	85.7%	5.4%
19	29.4%	69.1%	1.5%
TOTAL	51.9%	44.3%	4.0%

*B. Alignment of NTTP with Theory and Evidence on Effective Teacher PD*

The three content areas above not only attempt to address the needs of the education system in

China specifically, but are also in line with theory on what comprises effective teacher PD programs more generally. For example, empirical evidence from experimental studies (e.g., Carpenter et al. 1989), meta-analyses (e.g., Kennedy 1998), and case studies (e.g., Cohen 1990) focus on the importance of subject content knowledge and pedagogy. Peressini et al. (2004) and Desimone (2009) further provide conceptual frameworks highlighting the importance of teachers developing subject-specific knowledge, instructional skills, and a professional identity.

The additional delivery features (follow-up and evaluation) that were included in the evaluated program are also deemed effective in the theoretical literature: teachers may require a combination of incentives, evaluation and feedback to ensure they put what they learned in PD programs into practice (Guskey 1995). Teachers that go through PD programs may fail to implement what they learned in the programs due to insufficient follow-up (Cohen 1990; Lieberman 1995; Corcoran 1995; Guskey 1995; Schifter, Russell and Bastable 1999; 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. Moreover, even if teachers are able to acquire knowledge and skills from teacher PD programs, the programs may fail to hold trainees accountable for improving their teaching habits (Subirats and Nogales 1989; Braslavsky and Birgin 1992).

*Similarities between the NTTP and Teacher PD in other Developing Countries.*—The principles underlying the NTTP are also mirrored in teacher training programs across the developing world. Table A3 below compares the NTTP with 48 at-scale teacher PD programs in low- and middle-income countries using standardized indicators from the World Bank's In-Service Teacher Training Instrument (ITTSI) (Popova et al. 2018). In most respects, China's NTTP is similar to large-scale teacher training programs in other developing countries. Specifically, in terms of overarching program characteristics, it was designed by the government, involves some form of teacher evaluation, and provides complementary materials as part of the program. In terms of content, much like most programs, the NTTP has a focus on subject content or subject-specific pedagogy. With regards to program delivery, the NTTP has an emphasis on lectures and discussion that is typical across programs, and similarly involves a block of initial face-to-face training with few follow-up visits.

Three areas where the NTTP differs somewhat from other countries' programs are in the number of training days, whether training covers materials development, and the proportion of training

time spent on lectures versus spending time practicing with students. The NTTP offers no training on materials development because China uses standardized curricula across the country.

TABLE A3. COMPARING THE NTTP WITH GLOBAL TEACHER TRAINING PROGRAMS

	NTTP	Global Programs
<i>Overarching Aspects</i>		
Teachers pay some cost for the training (including their own transport)	No	52.08%
Program designed by Government	Yes	79.55%
Teachers evaluated	Yes	64.58%
Program includes complementary materials	Yes	95.83%
Program provides textbooks	No	29.17%
Program provides lesson plans/videos	Yes	54.17%
<i>Content</i>		
Primary focus of the training is subject content or subject-specific pedagogy	Yes	52.17%
Subject focus is math	Yes	54.17%
Training involves lectures	Yes	60.43%
Training involves discussion	Yes	84.17%
Training involves materials development	No	72.92%
Training involves training on how to conduct diagnostics	Yes	35.42%
Training involves lesson planning	Yes	62.50%
Training involves the use of scripted lessons	No	43.75%
<i>Delivery (for in-person training only)</i>		
Total hours of face-to-face training	120	13.26
Over how many weeks?	3	7.22
Proportion of training spent in lectures	73%	48.10%
Proportion of training spent practicing with students	0	15.57%
Program includes follow-up visits	Yes	49.64%
Number of follow-up visits	1	2.2
Follow-up visits to provide in-class pedagogical support	No	37.50%
Follow-up visits to review material	No	10.42%
Follow-up visits for monitoring	Yes	33.33%

*Notes:* Table compares the NTTP with 48 at-scale teacher PD programs in low- and middle-income countries using standardized indicators from the World Bank's In-Service Teacher Training Instrument (ITTSI) (Popova et al. 2018). All global data are means across the 48 programs. For binary variables, NTTP data take the form Yes/No, while global data report the percentage of programs for which Yes.

## Online Appendix B: Supplementary Tables

TABLE B1 – 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)
<i>Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)</i>								
(1)	PD	-0.030 (0.071)	0.088 (1.047)	0.006 (0.066)	0.094 (0.070)	0.010 (0.010)	0.257 (1.231)	-0.017 (0.069)
(2)	PD + Follow-up	-0.050 (0.071)	1.066 (1.001)	-0.036 (0.065)	0.106 (0.069)	-0.000 (0.014)	1.407 (1.188)	-0.092 (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
<i>Panel B: Comparing PD + Evaluation versus PD (left-out group)</i>								
(6)	PD + Evaluation	-0.042 (0.072)	0.164 (1.016)	0.042 (0.066)	0.055 (0.071)	0.011 (0.011)	-0.166 (1.229)	0.111 (0.067)
(7)	Observations	198	198	198	198	198	198	198
<i>Panel C: Comparing PD + Evaluation versus Control (left-out group)</i>								
(8)	PD + Evaluation	-0.059 (0.070)	0.752 (1.006)	0.001 (0.066)	0.131 (0.068)	0.010 (0.010)	0.867 (1.180)	0.004 (0.070)
(9)	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. We also conduct joint tests of the significance of the baseline covariates for each of the treatment comparisons. P-values from joint tests of the significance of baseline covariates across each treatment comparison follow in parentheses and all greater than 0.10: PD versus Control (0.742); PD + Follow-up versus Control (0.720); PD versus PD + Follow-up (0.808); PD + Evaluation versus PD (0.439); PD + Evaluation versus Control (0.578).

TABLE B2 – 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)
<i>Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)</i>							
(1)	PD	-0.016 (0.068)	-0.084 (0.042)	-0.002 (0.011)	0.025 (0.022)	0.014 (0.027)	-0.004 (0.075)
(2)	PD + Follow-up	-0.004 (0.072)	-0.032 (0.048)	-0.008 (0.011)	-0.011 (0.024)	-0.017 (0.029)	0.063 (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
<i>Panel B: Comparing PD + Evaluation versus PD (left-out group)</i>							
(6)	PD + Evaluation	0.075 (0.066)	-0.067 (0.043)	-0.014 (0.010)	0.008 (0.022)	0.015 (0.027)	-0.065 (0.071)
(7)	Observations	11,153	11,136	11,164	11,176	11,173	11,125
<i>Panel C: Comparing PD + Evaluation versus Control (left-out group)</i>							
(8)	PD + Evaluation	0.025 (0.069)	-0.091 (0.043)	-0.012 (0.011)	0.012 (0.022)	0.007 (0.027)	-0.001 (0.073)
(9)	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. We also conduct joint tests of the significance of the baseline covariates for each of the treatment comparisons. P-values from joint tests of the significance of baseline covariates across each treatment comparison follow in parentheses and all greater than 0.10: PD versus Control (0.231); PD + Follow-up versus Control (0.568); PD versus PD + Follow-up (0.579); PD + Evaluation versus PD (0.383); PD + Evaluation versus Control (0.191).

TABLE B3 – 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)
<i>Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)</i>						
(1) PD	-0.021 (0.067)	0.081 (0.040)	-0.004 (0.011)	0.021 (0.022)	0.009 (0.027)	-0.015 (0.076)
(2) PD + Follow-up	-0.001 (0.070)	-0.045 (0.046)	-0.009 (0.012)	-0.014 (0.024)	-0.016 (0.030)	0.056 (0.073)
(3) Endline attrition	-0.661 (0.075)	0.366 (0.054)	-0.147 (0.028)	-0.085 (0.026)	-0.060 (0.026)	-0.031 (0.068)
(4) PD # Endline attrition	0.047 (0.113)	-0.042 (0.086)	0.019 (0.036)	0.074 (0.038)	0.057 (0.041)	0.136 (0.094)
(5) PD + Follow-up # Endline attrition	-0.038 (0.105)	0.129 (0.071)	0.028 (0.038)	0.028 (0.039)	-0.055 (0.037)	0.102 (0.097)
(6) Observations	16,632	16,613	16,640	16,657	16,654	16,579
<i>Panel B: Comparing PD + Evaluation versus PD (left-out group)</i>						
(7) PD + Evaluation	0.082 (0.064)	-0.066 (0.041)	-0.011 (0.011)	0.012 (0.023)	0.020 (0.028)	-0.082 (0.073)
(8) Endline attrition	-0.660 (0.069)	0.442 (0.051)	-0.110 (0.028)	-0.037 (0.032)	-0.050 (0.031)	0.048 (0.060)
(9) PD + Evaluation # Endline attrition	-0.028 (0.111)	-0.042 (0.080)	-0.024 (0.036)	-0.001 (0.043)	-0.026 (0.044)	0.082 (0.098)
(10) Observations	11,153	11,136	11,164	11,176	11,173	11,125
<i>Panel C: Comparing PD + Evaluation versus Control (left-out group)</i>						
(11) PD + Evaluation	0.026 (0.068)	-0.095 (0.042)	-0.012 (0.011)	0.010 (0.022)	0.008 (0.028)	-0.018 (0.074)
(12) Endline attrition	-0.669 (0.075)	0.362 (0.054)	-0.148 (0.028)	-0.084 (0.026)	-0.057 (0.026)	-0.029 (0.068)
(13) PD + Evaluation # Endline attrition	0.008 (0.114)	0.029 (0.083)	0.016 (0.036)	0.045 (0.037)	-0.017 (0.039)	0.158 (0.100)
(14) 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.

TABLE B4 – TESTS FOR DIFFERENTIAL ATTRITION ACROSS TREATMENT AND CONTROL ARMS

		Midline	Endline
		(1)	(2)
<i>Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)</i>			
(1)	PD	-0.002 (0.006)	0.001 (0.005)
(2)	PD+Follow-up	0.005 (0.009)	0.006 (0.008)
(3)	Observations	15,296	
<i>Panel B: Comparing PD + Post-training Evaluation versus PD (left-out group)</i>			
(4)	Post-training Evaluation	-0.009 (0.007)	-0.002 (0.001)
(5)	Observations	11,180	
<i>Panel C: Comparing PD + Post-training Evaluation versus Control (left-out group)</i>			
(6)	PD + Post-training Evaluation	-0.003 (0.006)	0.005 (0.013)
(7)	Observations	11,403	

*Notes:* Table shows estimated coefficients and standard errors from a regression of a dummy variable indicating that a student was absent from the midline or endline survey on indicators for treatment and controlling for randomization strata. Cluster-robust SEs in parentheses. PD stands for professional development. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

TABLE B5 – IMPACTS ON STUDENT ACHIEVEMENT IN SPILLOVER SAMPLE (AT ENDLINE)

	(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel A: Comparing PD as well as PD + Follow-up versus Control (left-out group)</i>						
(1) PD	-0.033 (0.040)	-0.025 (0.034)				
(2) PD + Follow-up	-0.018 (0.041)	-0.010 (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				
<i>Panel B: Comparing PD + Evaluation versus PD (left-out group)</i>						
(6) PD + Evaluation			0.070 (0.038)	0.057 (0.037)		
(7) Observations			10,332	10,050		
<i>Panel C: Comparing PD Evaluation versus Control (left-out group)</i>						
(8) PD + Evaluation					0.007 (0.039)	0.018 (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.