Teacher Professional Development around the World: The Gap between Evidence and Practice

Anna Popova, David K. Evans, Mary E. Breeding, and Violeta Arancibia

Many teachers in low- and middle-income countries lack the skills to teach effectively, and professional development (PD) programs are the principal tool that governments use to upgrade those skills. At the same time, few PD programs are evaluated, and those that are evaluated show highly varying results. This paper proposes a set of indicators—the In-Service Teacher Training Survey Instrument—to standardize reporting on teacher PD programs. An application of the instrument to 33 rigorously evaluated PD programs shows that programs that link participation to career incentives, have a specific subject focus, incorporate lesson enactment in the training, and include initial face-to-face training tend to show higher student learning gains. In qualitative interviews, program implementers also report follow-up visits as among the most effective characteristics of their professional development programs. This paper then uses the instrument to present novel data on a sample of 139 government-funded, at-scale professional development programs across 14 countries. The attributes of most at-scale teacher professional development programs differ sharply from those of programs that evidence suggests are effective, with fewer incentives to participate in PD, fewer opportunities to practice new skills, and less follow-up once teachers return to their classrooms.

JEL Codes: I20, J24, O10
Keywords: Education quality, teacher training, professional development, economic development.

Good teachers have a major impact on student performance, both over the course of the school year (Araujo et al. 2016) and into adulthood (Chetty, Friedman, and Rockoff 2014). However, teachers in low- and middle-income countries often lack the skills they need to teach students effectively. Across seven African countries, only

doi: 10.1093/wbro/lkab006

seven percent of fourth-grade teachers had the minimum knowledge necessary to teach language; in four countries, the statistic was zero percent. For math teaching, 68 percent had the minimum knowledge needed to teach math—higher than the seven percent for language, but still leaving one in three teachers with insufficient knowledge. Teachers also scored woefully low in terms of pedagogical knowledge—their ability to prepare a lesson, formulate questions that would elicit student knowledge effectively, and their performance in the classroom (Bold et al. 2017).

The principal tool that countries across the income spectrum use to improve the knowledge and skills of their practicing teachers is professional development (PD), which refers to on-the-job training activities ranging from formal, lecture-style training to mentoring and coaching. However, few PD programs are rigorously evaluated, and among those that are, the evidence of their effectiveness is wildly mixed. Some programs are effective: training teachers to provide literacy instruction using students' mother tongue in Uganda and training teachers to evaluate student performance more regularly and adjust teaching based on those evaluations in Liberia both had sizeable impacts on student reading ability (Piper and Korda 2011; Kerwin and Thornton 2021). Others demonstrate opposite results: a large-scale, government-implemented PD program in China had zero impact on teacher knowledge, teaching practices, or student learning outcomes (Loyalka et al. 2019), and a program that trained teachers to engage their middle school math students more actively in learning in Costa Rica resulted in worse learning outcomes for students (Berlinski and Busso 2017). Indeed, there is much more variation in effectiveness across teacher training programs than across education programs more broadly (McEwan 2015; Evans and Popova 2016a). With this limited and highly variable evidence, policymakers and practitioners may be left puzzled as to how to structure teacher PD programs effectively.

In this paper, we propose a set of indicators—the In-service Teacher Training Survey Instrument, or ITTSI—to allow comparisons across teacher PD programs with varying impacts. On average, existing studies of PD programs only report on about half of these indicators. We supplement that information through interviews with implementors of evaluated PD programs. We compare the characteristics of 33 rigorously evaluated PD programs to identify which characteristics are associated with larger student learning gains. We then gather data from 139 government-funded, at-scale PD programs across 14 countries. Like most at-scale government programs, none of these programs have been evaluated rigorously. We compare the two samples to examine whether the PD programs that most teachers actually experience exhibit similar characteristics to those of PD programs that have been evaluated and shown to produce sizeable student learning gains.

When we apply our instrument to evaluated PD programs, results suggest that programs deliver high student learning gains when they link participation in PD to incentives such as promotion or salary implications, when they have a specific subject focus, when teachers practice enacting lessons during the training, and when training has at least an initial face-to-face aspect. Meanwhile, program implementers highlight two characteristics of effective training in interviews: mentoring follow-up visits after the PD training, and complementary materials such as structured lesson plans to help teachers apply what they have learned during PD.

When we subsequently use the ITTSI to characterize a sample of at-scale, government-funded PD programs around the world, we find a divergence in the characteristics common to these programs and those that typify evaluated programs that were found to be effective. Relative to top-performing PD programs—defined as those found to be the most effective at increasing student learning—very few at-scale PD programs are linked to any sort of career opportunities, such as promotion or salary implications. Similarly, in-school follow-up support and including time to practice with other teachers is less common among at-scale PD programs. This highlights a substantial gap between the kind of teacher PD supported by research and that currently being provided by many government-funded, at-scale programs.

These results have implications for both researchers and policymakers. For researchers, future evaluations will contribute much more to an understanding of how to improve teachers' skills if they report more details of the characteristics of the PD programs. Our proposed set of indicators can serve as a guide. For policymakers, at-scale PD programs should incorporate more aspects of successful, evaluated PD programs, such as incentives, practice, and follow-up in-school support. For both, more programs can be evaluated at scale, using government delivery systems, in order to improve the skills of teachers in the future.

Background

Conceptual Framework

The defining attributes of teacher professional development programs fall principally into three categories. The first is the content of the PD program: What is taught? The second is the delivery of the PD program: Who is teaching, when, and for how long? The third is the organization of the program beyond content and delivery: What are the scale and resources of the program? Are there incentives for participation? Was it designed based on a diagnostic of teachers? In this section, we discuss the theory behind each of these three categories.

On the content, PD programs focused on subject-specific pedagogy are likely to be most effective. General pedagogical knowledge—i.e., broad strategies of classroom management and organization—may contribute to student learning, driving the recent development of a range of classroom observation instruments (La Paro and Pianta 2003; Molina et al. 2018). However, different subjects require radically different pedagogies (Shulman 1986; Villegas-Reimers 2003). A highly scripted approach may work to teach early grade reading, whereas teaching science or civics in later grades—for example—may require more flexible approaches. PD programs that focus on arming teachers with subject-specific pedagogy are thus likely to make the largest contribution to student learning.

With respect to the delivery, the method, trainers, duration, and location of instruction all play a role. First, because working, professional teachers are the students in PD, principles of adult education are relevant to the method of instruction. Adult education tends to work best with clear applications rather than a theoretical focus (Cardemil 2001; Knowles, Holton, and Swanson 2005). The method of instruction should include concrete, realistic goals (Baker and Smith 1999) and the teaching of formative evaluation so that teachers can effectively evaluate their own progress towards their teaching goals (Bourgeois and Nizet 1997). Second, the quality of trainers—i.e., those providing the PD—is crucial to learning (Knowles, Holton, and Swanson 2005). In terms of the delivery of PD, this calls into question the common cascade model of PD in low-income environments, in which both information and pedagogical ability may be diluted as a master trainer trains another individual as a trainer, who may go on to train another trainer below her, and so forth.

Third, on the duration of instruction, there is no theoretical consensus on exactly how long training should last, although there is suggestive empirical evidence in the literature in favor of sustained contact over a significant period of time and against brief, one-time workshops (Desimone 2009). Fourth, on the location of instruction, teacher PD in the school ("embedded") is likely to be most effective so that participating teachers can raise concrete problems that they face in the local environment, and they can also receive feedback on actual teaching (Wood and McQuarrie 1999). However, this will depend on the environment. In very difficult teaching environments, some degree of training outside the school may facilitate focus on the part of the trainees (Kraft and Papay 2014).

Finally, the organization of the PD—which includes overarching aspects such as who is organizing it, for whom, and how—provides an important backdrop when we consider any PD program. This includes aspects such as the scale, cost, and targeting of the program. In general, it is predictably easier to provide high-quality PD through smaller scale, higher cost programs that provide more tailored attention to a given teacher. In terms of targeting, teacher PD will work best if it adjusts at different points in the teachers' careers: One would not effectively teach a brand-new teacher in the same way as one would train a teacher with 20 years of experience (Huberman 1989). Teachers see their greatest natural improvements in the first five years of teaching, which may be an indicator of greater skill plasticity, so there may be benefits to leveraging that time (TNTP 2015).

What Works in High-Income Countries?

A full review of the literature in high-income countries is beyond the scope of this study. However, it may be useful to highlight recent work on in-service teacher PD from the United States—which spends almost $18,000 per teacher and 19 days of teacher time on training each year (TNTP 2015)—and other high-income countries, in order to ensure that low- and middle-income countries are not ignoring well-established evidence. Several promising themes that emerge from this work are the importance of making PD specific and practical, providing sustained follow-up support for teachers, and embedding it in the curriculum.

Specific and practical teacher PD finds support from multiple reviews of teacher PD studies in high-income countries, which conclude that concrete, classroom-based programs make the most difference to teachers (Darling-Hammond et al. 2009; Walter and Briggs 2012). More recently, a meta-analysis of 196 randomized evaluations of education interventions—not just PD—in the United States that measure student test scores as an outcome examined the impact of both "general" and "managed" professional development, relative to other interventions (Fryer 2017). General PD may focus on classroom management or increasing the rigor of teachers' knowledge, whereas managed professional development prescribes a specific method, with detailed instructions on implementation and follow-up support. On average, managed PD increased student test scores by 2.5 times (0.052 standard deviations) as much as general PD and was at least as effective as the combined average of all school-based interventions. A recent review of nearly 2,000 impact estimates from 747 randomized controlled trials of education interventions in the United States proposes that an effect size of 0.05 be considered a "medium" effect size, higher than the average effect size, weighted by study sample size (Kraft 2020), which suggests that these are not trivial impacts.

The importance of sustained follow-up support is echoed by another U.S.-focused review, which found that PD programs with significant contact hours (between 30 and 100 in total) over the course of six to twelve months were more effective at raising student test scores (Yoon et al. 2007). Likewise, a narrative review of U.S. studies concluded that the most effective programs are not "one-shot workshops": they are sustained, intense, and embedded in the curriculum (Darling-Hammond et al. 2009).

Despite these conclusions, the experimental or quasi-experimental evidence is thin, even in high-income countries. The meta-analysis of 196 evaluations of education interventions included just nine PD studies (Fryer 2017), and another review of 1,300 PD studies identified just nine that had pre- and post-test data and some sort of control group (Yoon et al. 2007). Similarly, a review of PD in mathematics found more than 600 studies of math PD interventions, but only 32 used any research design to measure effectiveness, and only five of those were high-quality randomized trials (Gersten et al. 2014). The question of what drives effective teacher PD remains understudied, even in high-income environments.

We expect teachers in lower and middle-income countries to learn in fundamentally similar ways to their high-income counterparts. However, lower resource contexts are typically characterized by more binding cost constraints and lower teacher and coach pedagogical capacity. These challenges may make certain elements of PD programs more and less relevant in lower-income contexts. Teachers and coaches in low- and middle-income countries may benefit from more prescriptive instructions on implementation and, while they too require ongoing follow-up as part of PD, this may need to be provided in lower-cost forms, whether in group sessions, using technology for remote coaching, or training school principals and experienced peer teachers as coaches.

Methods

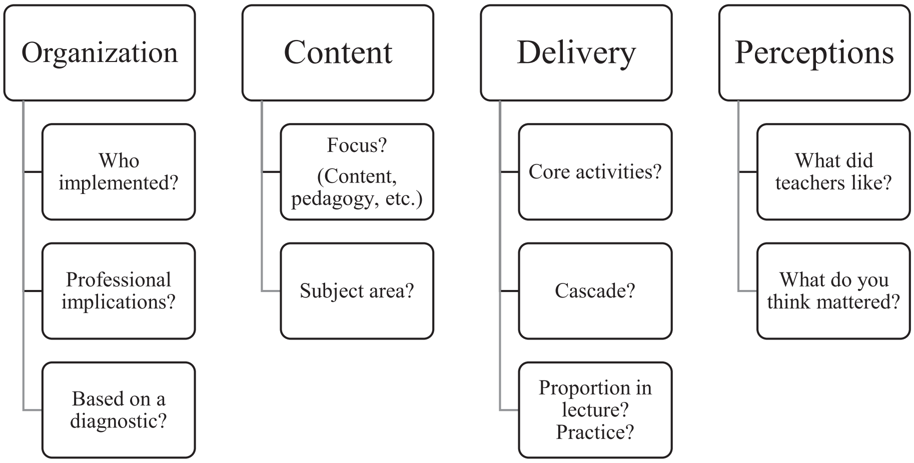
To understand which characteristics of PD programs are associated with student test score gains, and to analyze the degree to which these effective characteristics are incorporated into at-scale PD programs in practice, we first developed a standardized instrument to characterize in-service teacher training. Second, we applied this instrument to already evaluated PD programs to understand which PD characteristics are associated with student learning gains. Third, we applied the survey instrument to a sample of at-scale PD programs to see how these programs line up with what the evidence suggests works in teacher training. The information we present thus comes from two different samples of PD programs: One sample of *evaluated* PD programs, those with impact evaluations that include student assessment results; and one sample of *at-scale*, government-funded PD programs.[1] The remainder of this section introduces the instrument briefly before describing its application to each of the two samples.

The In-Service Teacher Training Survey Instrument (ITTSI)

The ITTSI was designed based on the conceptual framework and empirical literature characterized in the previous sections, as well as on the authors' prior experience studying in-service teacher PD. We drafted an initial list of 51 key indicators to capture details about a range of program characteristics falling into three main categories: Organization, Content, and Delivery, paralleling the three elements of our conceptual framework (fig. 1). We supplement those categories with a fourth category, Perceptions, which we added to collect qualitative data from program implementors.

Taking each of these in turn, the Organization section includes items such as the type of organization responsible for the design and implementation of a given teacher training program, to whom the program is targeted, what (if any) complementary

Figure 1. Summary of the In-Service Teacher Training Survey Instrument (ITTSI)



Source: Authors' summary of the elements of the In-Service Teacher Training Survey Instrument, as detailed in supplementary online appendices A1 and A2.

materials it provides, the scale of the program, and its cost. The Content section includes indicators capturing the type of knowledge or skills that a given program aims to build among beneficiary teachers, such as whether the program focuses on subject content (and if so, which subject), pedagogy, new technology, classroom management, counseling, assessment, or some combination.

Delivery focuses on indicators capturing program implementation details, such as whether it is delivered through a cascade model, the profile of the trainers who directly train the teachers, the location of the training, the size of the sessions, and the time division between lectures, practice, and other activities. Finally, the Perceptions section includes indicators capturing program implementers' own perceptions of which elements were responsible for any positive impacts and which were popular or unpopular among teachers. We piloted the draft instrument by using it to collect data on a sample of evaluated programs, and validated its ability to accurately characterize the details of PD programs by sharing our results with a series of expert researchers and practitioners in teacher PD. We updated the indicators in light of this feedback, resulting in a final version of the instrument, which includes 70 indicators plus three pieces of metadata. Further information on the instrument can be found in the supplementary online appendices: Appendix A1 provides a more detailed description of instrument development; appendix A2 presents the final instrument (ITTSI); and appendix A3 presents the Brief In-Service Teacher Training Instrument (BITTSI), a supplementary instrument we developed containing a subset of the 13 most critical questions from the ITTSI based on our reading of the literature.

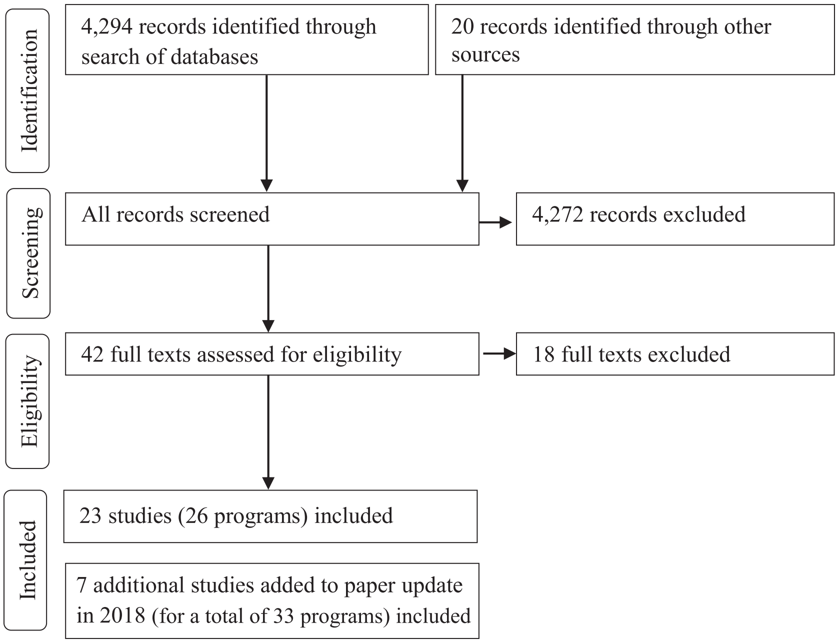
The ITTSI does not collect extensive data about the broader educational context. Context includes teacher policies (e.g., pre-service training and the structure of the teacher career), other education policies, and the current state of education (e.g., learning and absenteeism rates). Context matters for the impact of teacher PD programs. As a simple example, in a setting where student absenteeism is extremely high, teacher PD programs may have a limited impact on student learning due to few hours of contact between teachers and students. That said, certain principles of teacher PD may translate across cultures, even if the applications vary. Professionals need practice to master skills across contexts, so giving teachers the opportunity to practice lessons during training may be valuable across contexts, even if how they do that may vary. Other survey instruments have been developed and tested broadly to gather a wide range of data on the education system, notably the World Bank's Systems Approach for Better Education Results (SABER) (Rogers and Demas 2013). For a rich view of teacher PD in context, the ITTSI could be complemented with the SABER instrument or other data about the education system.

Applying the ITTSI to Evaluated PD Programs

We searched the existing literature on in-service teacher PD in low- and middle-income countries to identify a sample of PD programs that had been evaluated for their impact on student learning. Our inclusion criteria for the search were impact evaluations of primary and secondary education interventions in low- and middle-income countries that (a) focused primarily on in-service teacher PD or included this as a major component of a broader program, and (b) reported impacts of the program on student test scores in math, language, or science. We included both published and unpublished papers and did not restrict by year of authorship.

In order to identify papers fulfilling the above criteria, we searched a range of databases in 2016.[2] The search yielded 6,049 results and automatically refined the results by removing exact duplicates from the original results, which reduced the number of results to 4,294. To this we added 20 impact evaluations which mention teacher PD from a recent review (Evans and Popova 2016a). We examined the 4,314 results from both sources to exclude articles that—from their title and abstract—were clearly not impact evaluations of teacher training programs. This review process excluded 4,272 results and left 42 full articles to be assessed for eligibility. After going through the full texts, another 18 papers were excluded as the full text revealed that they did not meet the inclusion criteria. This yielded 23 papers, which evaluated 26 different PD programs. In February 2018, we updated this original sample with full articles published between 2016 and 2018 which fit the inclusion criteria. This resulted in seven new papers and teacher PD programs for a total of 30 papers evaluating 33 programs. The search process is detailed in fig. 2. The 30 papers are listed in supplementary online appendix A4.

Figure 2. Search Process and Results for Evaluated Professional Development Programs


Identification
Screening
Eligibility
Included
4,294 records identified through search of databases
20 records identified through other sources
All records screened
4,272 records excluded
42 full texts assessed for eligibility
18 full texts excluded
23 studies (26 programs) included
7 additional studies added to paper update in 2018 (for a total of 33 programs) included

Source: Constructed by the authors based on the search described in the text.
Note: The 30 papers documenting the evaluation of the final 33 programs are listed in supplementary online appendix A4.

Data collection and coding for the sample of 33 evaluated programs comprised two phases. The first of these phases consisted of carefully reviewing the impact evaluation studies and coding the information they provided. The draft version of the instrument for which we collected data included 51 indicators in total, and on average, information on 26 (51 percent) of these indicators was reported in the impact evaluations. Crucially, the amount of program information reported across the impact evaluations varies noticeably by topic (table 1). Sixty-four percent of details concerning the organization of teacher training programs—such as whether the program was designed by a government or by a non-governmental organization (NGO)—can be extracted from the evaluations. In contrast, on average, only 47 percent of information concerning program content and 42 percent of information concerning program delivery is reported.

The second phase of data collection sought to fill this gap in reported data by interviewing individuals involved in the actual implementation of each program. To do this, we emailed the authors of each of the impact evaluations in our sample, asking them to connect us with the program implementers. After three attempts to contact the implementers, we received responses from authors for 25 of the 33

Table 1. Data Available on Evaluated Programs from Studies vs. Interviews

	Percentage data collected		
	From impact evaluation reports only	After interviews with implementers	Total number of indicators
Organization	64%	78%	27
Content	47%	66%	10
Delivery	42%	69%	14
TOTAL	51%	75%	51
For interviewed programs only		98%	51

Source: Constructed by the authors based on the application of the In-Service Teacher Training Survey Instrument items (supplementary online appendix A2) to the 33 professional development programs identified (supplementary online appendix A4).

Note: Percentage data collected refers to the percentage of indicators for which data were collected across the 33 programs in our evaluated sample. This is calculated by the number of programs for which each indicator has data, summed for every indicator in a given section (or total) and divided by the number of indicators in that section (or total), and finally divided by the 33 programs.

programs. We contacted all of the individuals to whom the authors referred us—who in many cases directed us to more relevant counterparts—and were eventually able to hold interviews with program implementers for 18 of the 33 programs.[3] The interviews loosely followed the survey instrument, but included open-ended questions and space for program implementers to provide any additional program information that they perceived as important.

The ITTSI data were gathered retrospectively for this study, which means that in most cases, the evaluation results (and so whether or not the program was effective) were likely to have been known to the interviewee. We propose three reasons that this should not pose a substantive problem for the quality of the data. First, most of the indicators have no normative response. Whether a program is government- or researcher-designed or implemented, whether it has a subject focus or a general pedagogy focus, or whether or not it has a distance learning element have no obvious "right" answers. Second, the survey was administered to program implementers, who usually were not part of the team of researchers who evaluated the program, so they had little stake in confirming research results. Third, the survey had low stakes: interviewees knew that we were independent researchers doing a synthesis review. In some cases, the PD program being discussed no longer existed in the same form. For future PD studies, these data could be collected at the design stage of programs.

For the 18 programs for which we conducted interviews, we were able to collect information for an average of 50 out of the 51 (98 percent) indicators of interest. Consequently, conducting interviews decreased the differences in data availability across categories. The pooled average of indicators for which we had information after conducting interviews (for interviewed and not interviewed programs combined) increased to 79 percent for Organization indicators, 68 percent of Content indicators, and 72 percent of Delivery indicators (table 1).

For our sample of evaluated in-service teacher PD programs, we analyze which characteristics of teacher training programs are associated with the largest improvements in student learning, as measured by test score gains. We conduct both quantitative and qualitative analyses. The analytical strategy for the quantitative analysis essentially consists of comparing means of student learning gains for programs with and without key characteristics, using a bivariate linear regression to derive the magnitude and statistical significance of differences in means. We do not carry out multivariate regression analysis because of the small sample; thus, these results are only suggestive, as multiple characteristics of programs may be correlated. Because we are testing each coefficient separately, we are not able to test the relative value of coefficients, so differences in point estimates are only suggestive.

In preparation for this analysis, we standardize the impact estimates for each of the programs. We convert the program characteristic variables to indicator variables wherever possible to facilitate comparability of coefficients. Although our sample of impact evaluations has a common outcome—impact on student test scores—these are reported on different scales across studies, based on different sample sizes.[4] We standardize these effects and the associated standard errors in order to be able to compare them directly. Supplementary online appendix A5 provides mathematical details of the standardization.

Turning to the independent variables, as originally coded, the 51 indicators for which we collected information capturing various design and implementation characteristics of the PD programs took a number of forms. These consisted of indicator variables (e.g., the intervention provides textbooks alongside training = 0 or 1), categorical variables (e.g., the primary focus of the training was subject content [= 1], pedagogy [= 2], new technology [= 3]), continuous variables (e.g., the proportion of training hours spent practicing with students), and string variables capturing open-ended perceptions (e.g., which program elements do you think were most effective?). To maximize the comparability of output from our regression analysis we convert all categorical and continuous variables into indicator variables.[5]

We then conduct our bivariate regressions on this set of complete indicator variables with continuous impact estimates on test scores as the outcome variable for each regression. Because of the limitations associated with running a series of bivariate regressions on a relatively small sample of evaluations, we propose the following robustness check. First, we estimate robust Eicker-Huber-White (EHW) standard errors as our default standard errors (reported in tables 2–4) and assess significance according to p-values associated with these. Second, we estimate bootstrapped standard errors and the associated p-values. Third, we run Fisher randomization tests to calculate exact p-values, a common approach in the context of small samples.[6]

Table 2. Organization – Bivariate Regressions with Robustness Checks

Organization	Coefficient	Standard error	Significant	Programs with characteristic	Total programs	Robust
Designed by government	0.068	0.079		5	33	
Designed by NGO or social enterprise	0.012	0.062		13	33	
Designed by researchers	−0.036	0.067		14	33	
Implemented by Government	−0.016	0.062		9	33	
Implemented by NGO or social enterprise	0.012	0.062		13	33	
Implemented by researchers	0.001	0.078		11	33	
Design not based on diagnostic	0.041	0.099		4	33	
Design based on informal diagnostic	−0.002	0.062		8	33	
Design based on formal diagnostic	0.007	0.080		11	33	
Targeting by geography	0.017	0.063		16	30	
Targeting by subject	−0.065	0.057		9	30	
Targeting by grade	−0.040	0.058		25	31	
Targeting by years of experience	0.101	0.051	*§	2	30	X
Targeting by skill gaps	−0.060	0.034	*§	1	30	
Targeting by contract teachers	0.044	0.075		3	30	
Participation has no implications for status, salary or promotion	−0.120	0.056	**§†	12	33	X
Participation has status implications only	0.004	0.071		2	33	
Participation has implications for salary or promotion	0.023	0.056		10	33	
Teachers are not evaluated	−0.084	0.073		7	33	
Positive consequence if teachers are well evaluated	0.025	0.062		4	33	
Negative consequence if teachers are poorly evaluated	0.054	0.075		2	33	
Program provides materials	0.051	0.069		26	30	
Program provides textbooks	0.081	0.123		6	28	
Program provides storybooks	0.106	0.087		9	28	
Program provides computers	−0.029	0.086		4	28	
Program provides teacher manuals	−0.056	0.063		16	29	
Program provides lesson plans/videos	−0.006	0.097		9	28	
Program provides scripted lessons	−0.030	0.073		7	29	
Program provides craft materials	−0.061	0.039		3	28	
Program provides other reading materials (flashcards, word banks, reading pamphlets)	0.132	0.080		10	28	
Program provides software	−0.026	0.061		8	29	
Number of teachers trained > median (= 110)	−0.012	0.065		9	19	
Number of schools in program > median (= 54)	0.091	0.066		14	28	
Program age (years) > median (= 2)	0.057	0.075		8	25	
Dropouts in last year	0.083	0.071		8	15	

Source: Constructed by the authors based on data extracted from 33 professional development programs (supplementary online appendix A4) using the In-Service Teacher Training Survey Instrument, and analyzed by regression, as described in the text.

Note: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$ correspond to the significance of *p*-values of robust standard Noteerrors. § corresponds to significance at the 10 percent level or higher for bootstrapped standard errors. † corresponds to significance at the 10 percent level or higher for the Fisher Randomization tests. Numbers specified in parentheses in variable labels are the reported medians for dummy variables in which the variable equals 1 if greater than the median. Total programs refers to the number of programs that report whether or not they have the characteristic. The robust column includes an X if the finding is statistically significant across at least two methods and if the finding is driven by two or more evaluations (i.e., not a single evaluation).

Table 3. Content – Bivariate Regressions with Robustness Checks

Content	Coefficient	Standard error	Significant	Programs with characteristic	Total programs	Robust
Focus is subject content	0.099	0.060		21	33	
Focus is pedagogy	0.078	0.060		19	33	
Focus is technology	0.060	0.056		7	33	
Focus is counseling	−0.199	0.056	***§†	3	33	X
Focus is classroom management	−0.020	0.116		4	33	
Focus is a specific tool	−0.118	0.038	***§	3	33	X
No subject focus	−0.236	0.054	***§†	2	33	X
Subject focus is literacy/language	0.069	0.062		17	33	
Subject focus is math	−0.086	0.058		5	33	
Subject focus is science	−0.038	0.049		3	33	
Subject focus is information technology	0.086	0.033	**§	1	33	
Subject focus is language & math	0.023	0.095		2	33	
Subject focus is other	−0.103	0.033	***§	1	33	
Training involves lectures	0.020	0.031		19	20	
Training involves discussion	0.004	0.080		15	20	
Training involves lesson enactment	0.102	0.055	*§†	12	20	X
Training involves materials development	0.010	0.055		4	20	
Training involves how to conduct diagnostics	0.070	0.079		5	21	
Training involves lesson planning	0.061	0.083		12	25	
Training involves use of scripted lessons	0.018	0.111		8	24	

Source: Constructed by the authors based on data extracted from 33 professional development programs (supplementary online appendix A4) using the In-Service Teacher Training Survey Instrument, and analyzed by regression, as described in the text.

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ correspond to the significance of p-values of robust standard errors. § corresponds to significance at the 10 percent level or higher for bootstrapped standard errors. †corresponds to significance at the 10 percent level or higher for the Fisher Randomization tests. Total programs refers to the number of programs that report whether or not they have the characteristic. The robust column includes an X if the finding is statistically significant across at least two methods and if the finding is driven by two or more evaluations (i.e., not a single evaluation).

We report significance under each of these methods separately and report results as robust if they are significant under at least two of the three methods, and if the significant effect is driven by at least two observations—i.e., the results are not explained by a single PD program.

We supplement this regression analysis with a qualitative analysis of what works, relying on the self-reported perceptions of program implementers along three dimensions: (a) Which program elements they identified as most responsible for any positive impacts on student learning; (b) which elements, if any, teachers particularly liked; and (c) which elements, if any, teachers particularly disliked.

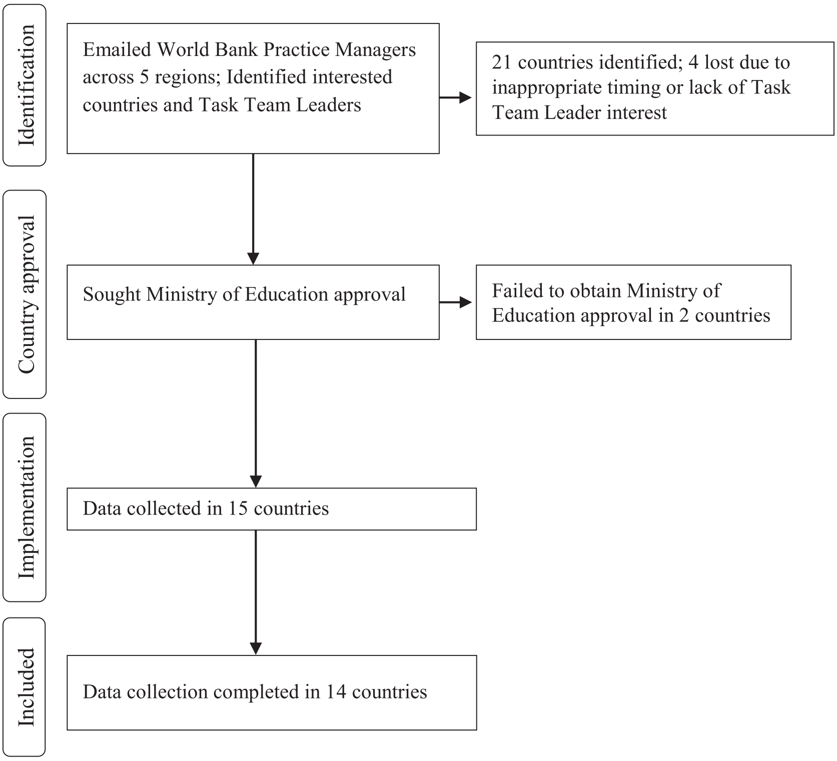
Table 4. Delivery – Bivariate Regressions with Robustness Checks

Delivery	Coefficient	Standard error	Significant	Programs with characteristic	Total programs	Robust
Cascade training model	−0.026	0.073		14	27	
Trainers are primary or secondary teachers	0.005	0.069		5	33	
Trainers are experts - university professors/graduate degrees in education	−0.048	0.118		7	33	
Trainers are researchers	−0.042	0.049		3	33	
Trainers are local government officials	−0.019	0.052		8	33	
Trainers are education university students	0.148	0.032	***§	1	33	
Initial period of face-to-face training for several days in a row	0.140	0.041	***§	30	32	X
Total hours of face-to-face training > median (= 48)	0.051	0.067		15	31	
Proportion of face-to-face training spent in lectures > median (= 50%)	−0.095	0.060		6	17	
Proportion of face-to-face training spent practicing with students > median (= 0)	0.058	0.054		7	19	
Proportion of face-to-face training spent practicing with teachers > median (33%)	0.155	0.094	†	9	19	
Duration of program (weeks) > median (= 2.5)	−0.038	0.068		15	30	
Training held at schools	−0.043	0.033		1	33	
Training held at central location including hotel conference room etc.	−0.126	0.064	*§†	19	33	X
Training held at university or training center	0.263	0.174	†	3	33	
Number of teachers per training session > median (= 26)	0.086	0.059		8	17	
Includes follow-up visits	0.108	0.070		19	25	
Follow-up visits for in-class pedagogical support	0.100	0.078		11	33	
Follow-up visits for monitoring	−0.022	0.052		8	33	
Follow-up visits to review material	0.139	0.112		3	33	
Includes distance learning	−0.100	0.050	*§	4	24	X
Duration of distance learning (months) > median (= 26)	−0.094	0.061		10	27	

Source: Constructed by the authors based on data extracted from 33 professional development programs (supplementary online appendix A4) using the In-Service Teacher Training Survey Instrument, and analyzed by regression, as described in the text.

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ correspond to the significance of p-values of robust standard errors. § corresponds to significance at the 10 percent level or higher for bootstrapped standard errors. † corresponds to significance at the 10 percent level or higher for the Fisher Randomization tests. Numbers specified in parentheses in variable labels are the reported medians for dummy variables in which the variable equals 1 if greater than the median. Total programs refers to the number of programs that report whether or not they have the characteristic. The robust column includes an X if the finding is statistically significant across at least two methods and if the finding is driven by two or more evaluations (i.e., not a single evaluation).

Figure 3. Sampling Process for At-Scale Professional Development Programs


Identification
Emailed World Bank Practice Managers across 5 regions; Identified interested countries and Task Team Leaders
21 countries identified; 4 lost due to inappropriate timing or lack of Task Team Leader interest
Country approval
Sought Ministry of Education approval
Failed to obtain Ministry of Education approval in 2 countries
Implementation
Data collected in 15 countries
Included
Data collection completed in 14 countries

Source: Constructed by the authors to reflect the process to identify at-scale professional development programs, as described in the text.

Applying the ITTSI to At-Scale PD Programs

The sampling process for at-scale programs is detailed in fig. 3. To obtain a sample of at-scale, government-funded PD programs across the world, we first identified four to five countries in each region where the World Bank has operations.[7] We worked with regional education managers at the World Bank in each region to select countries in which government counterparts and World Bank country teams had an interest in learning more about in-service teacher PD programs. We made clear that the exercise was appropriate for countries with any level of teacher PD, not specific to countries with recent reforms or innovations. The final set of countries sampled included Burkina Faso, Cambodia, El Salvador, The Gambia, Guinea, India (Bihar state), Jordan, Kazakhstan, the Kyrgyz Republic, Mauritania, Mexico (Guanajato, Oaxaca, and Puebla, and a national PD program for middle school teachers), Moldova, Niger, and the Russian Federation.

We then obtained permission from the Ministry of Education (MoE) or other relevant government counterparts in each country and worked with them to complete a roster, or listing, of all teacher PD programs conducted between 2012 and 2016.[8] The roster, available in supplementary online appendix A6, was created along with the ITTSI instrument and collects the following information about each of the teacher PD programs that received government funding: program name; program coordinator's name and contact information; the number of teachers trained; and the types of teachers targeted (e.g., pre-primary, primary, or secondary school teachers). In some countries, such as Mexico and India, where policymaking about teacher PD happens at the state level, we worked with individual states.

After receiving completed roster information about teacher PD programs in a country/state, we used the roster to select a sample of teacher PD programs to interview. In each country/state, we chose the sample by selecting the 10 largest teacher PD programs in terms of teacher coverage, defined as the number of teachers reached by the program during its most recent year of implementation. Of the 10 sampled programs for each country/state, the full ITTSI was administered to the two largest programs targeting primary school teachers and the largest program that targeted secondary school teachers. The brief version of the instrument, the BITTSI, was administered in the remaining seven programs in the country/state. In total, 48 at-scale programs completed the ITTSI and 91 at-scale programs completed the BITTSI across 14 countries.

We applied the ITTSI survey through a combination of phone interviews with and online surveys of PD program coordinators. In a few instances (in The Gambia, El Salvador, and Mexico), depending on the preferences of the program coordinator and their primary language, program coordinators were given the option of completing the ITTSI questionnaire online. For the majority of programs, however, we held phone interviews with program coordinators, in which we asked them the questions included in the ITTSI survey items directly and filled out the instrument ourselves with their responses.

The ITTSI survey applied to the sample of at-scale programs consists of 70 indicators. We were able to collect information for an average of 66 of the 70 (94 percent) indicators of interest for the 48 at-scale teacher PD programs to which the full ITTSI survey was applied, and for 26.5 of the 27 (97 percent) indicators—derived from 13 questions—for the 91 programs to which the BITTSI was applied.

For the sample of at-scale PD programs, we compare the average of observed characteristics of at-scale teacher PD programs with the average for evaluated PD programs that resulted in the largest improvements in student learning ("top performers"), as measured by student test score gains. To determine the characteristics of "top performers," we ranked all evaluated programs, using their standardized impact on student test scores. We then selected the top half of programs (16 programs, all of which displayed positive impacts), and calculated the average value of program indicators for those "top performers." We compare them to the means of at-scale PD programs in order to better understand the gap between at-scale PD practices and the best practices of top-performing PD programs.

Results

This section characterizes the specific characteristics of teacher PD programs that successfully improve student learning in low- and middle-income countries and how common these characteristics are across at-scale, government-funded programs. First, we present the results of our quantitative and qualitative analyses examining which PD characteristics are associated with large gains in student learning for the sample of evaluated programs. Second, we present descriptive statistics from the sample of at-scale PD programs and from the top-performing PD programs in the evaluated sample to shed light on how they differ in terms of those PD characteristics found to be associated with positive impacts on student learning.

Which PD Characteristics are Most Associated with Student Learning Among Evaluated Programs?

We discuss, for each of our categories—Organization, Content, and Delivery—those characteristics we observe to be most associated with student learning gains. Tables 2–4 present the results of our bivariate regressions for each of these categories in turn. In each case, we report the results with the three different methods of calculating significance as well as an indicator of robustness.

Among Organization (table 2), two characteristics are robustly associated with significant gains in student learning. These include linking career opportunities (improved status, promotion, or salary) to PD programs and targeting training programs based on teachers' years of experience. First, in teacher PD programs where participation has no implications for promotion, salary, or status increases, student learning is 0.12 standard deviations lower (significant at 95 percent). In other words, programs that do link participation to career incentives have higher effectiveness.[9] Second, targeting participant teachers by their years of experience is associated with 0.10 standard deviations higher student learning (significant at 90 percent). This is driven by two programs: the Balsakhi program in rural India, which trains women from the local community who have completed secondary school to provide remedial education to students falling behind (Banerjee et al. 2007); and the Science teacher training program in Argentina, which trains teachers in different structured curricula and coaching techniques and finds that coaching is only effective for less experienced teachers (Albornoz et al. 2018). Indeed, these are the only two programs out of the 33 that explicitly targeted teachers based on their experience, both of which resulted in student learning gains. In addition, the provision of complementary materials such as storybooks and other reading materials (e.g., flashcards or word banks) have large coefficients associated with improving student learning (0.11 and 0.13 standard deviations), although these are not statistically significant.

Among the Content variables (table 3), programs with a specific subject focus result in higher learning gains than more general programs. Specifically, programs with no subject focus show 0.24 standard deviations lower impact on student learning (significant at 99 percent). A deeper look reveals that within focus areas, programs that are not focused on a given academic subject—such as those focused on counseling—are associated with 0.2 lower standard deviations in student learning (significant at 99 percent). Lastly, when a teacher PD program involves teaching practice through lesson enactment, it is associated with a 0.10 standard deviation increase in student learning (significant at 90 percent).

Turning to Delivery characteristics (table 4), three characteristics of teacher PD programs are robust. First, teacher PD programs that provide consecutive days of face-to-face teacher training are associated with a 0.14 standard deviation increase in student learning (significant at 99 percent). Second, holding face-to-face training at a central location—such as a hotel or government administrative building (as opposed to a university or training center, which was the omitted category)—is associated with a 0.13 lower standard deviation in student learning (significant at 90 percent). Third, teacher PD training programs that are conducted remotely using distance learning are associated with a 0.10 standard deviation decrease in student learning (significant at 90 percent). In alignment with recent literature highlighting the overly theoretical nature of many training programs as an explanation for their limited effects on student learning—as well as the above finding that training programs that involve teaching practice are associated with 0.16 larger gains in student learning—the proportion of training time spent practicing with other teachers is highly correlated with learning impacts (although not consistently statistically significant). Also, the inclusion of follow-up visits to review material taught in the initial training—as opposed to visits for monitoring purposes alone or no follow-up visits—is associated with a 0.14 standard deviation higher program impact on student learning (not significant, but one of the largest coefficients). These findings support the literature that subject-focused teacher PD programs with consecutive days of face-to-face training that include time for teachers to practice with one another, are associated with improved student learning outcomes.

We supplement the quantitative results with an analysis of self-reported perceptions by the implementers of the evaluated programs. These concern the characteristics of their programs which they believe are most responsible for any positive effects on student learning, as well as those elements which were popular and unpopular among the beneficiary teachers. We elicited these perceptions using open-ended questions and then tallied the number of program implementers that mentioned a given program element in their response, albeit not necessarily using the exact same language as other respondents. These responses come from 18 interviewees, so they should be taken as suggestive. That said, the results broadly align with the quantitative results: Five of 18 interviewees—tied for the most common response—mentioned that mentoring follow-up visits were a crucial component in making their training work. Similarly, five of the 18 interviewees discuss the importance of having complementary materials, such as structured lessons or scripted materials that provide useful references in the classroom and help to guide teachers during the training sessions. The next most commonly reported elements were engaging teachers for their opinions and ideas—either through discussion or text messages—and designing the program in response to local context, building on what teachers already do and linking to everyday experiences: both were mentioned by four of 18 interviewees.

We also asked the program implementers about the program characteristics that they believed teachers liked and disliked the most about their training programs and, interestingly, we only found two common responses for what teachers particularly liked and one common response for what they disliked.[10] Seven of the 18 interviewees reported that the part of their program that teachers most enjoyed was that it was fun and engaging (or some variation of that). In other words, teachers appreciated that certain programs were interactive and involved participation and discussion rather than passive learning. In addition to having "fun" teacher PD programs, five of the 18 interviewees suggested that teachers especially liked the program materials provided to them. Similarly, in terms of unpopular program elements, four of the 18 program implementers we interviewed reported that teachers disliked the amount of time taken by participating in the training programs, which they perceived as excessive.

What Do We Learn from At-Scale PD Programs?

Government-funded, at-scale teacher PD programs have a number of characteristics in common (supplementary online appendix tables A7.1–A7.3). The vast majority are designed by government (80 percent) and implemented by government (90 percent). Almost all provide materials to accompany the PD (96 percent), and most include at least some lesson enactment (73 percent) and development of materials (73 percent). Most have a subject focus (92 percent) and include an initial period of face-to-face training for several days (85 percent). Most do not formally target teachers by subject (only 19 percent do), grade (31 percent), or years of experience (13 percent), and few have negative consequences if teachers are poorly evaluated (17 percent). These at-scale programs differ sharply from programs that are evaluated in general, as well as from top-performing evaluated programs specifically. We provide a full list of average characteristics of at-scale programs and all evaluated programs (not just top-performers) in supplementary online appendix tables A7.1–A7.3.

Our principal focus in this section is how at-scale programs compare to evaluated programs that deliver relatively high gains in student learning. We assess the top half of programs (N = 16) from the sample of evaluated programs by selecting those characteristics that produced the largest standard deviation increases in student assessment scores. In table 5, we compare the means of at-scale programs and top-performing, evaluated programs. We focus specifically on the characteristics shown to have a statistically significant relationship with student learning outcomes and those with large coefficients, identified for interest (as identified in tables 2–4).

Regarding Organization (table 5), two key characteristics—whether or not the training is linked to career opportunities and whether or not the program targets teachers based on their years of experience—are robustly associated with improved student learning gains. There are notable and substantive differences between top-performing PD programs and the sample of at-scale PD programs when it comes to providing incentives; 88 percent of top-performing PD programs link training to status or to new career opportunities such as promotion or salary, as compared to only 55 percent of at-scale programs. Our results suggest that without incentives, training may not have a meaningful impact. Furthermore, top-performing programs and at-scale PD programs are similar in the degree to which they target teachers based on their years of experience. For instance, 13.3 percent of top-performers and 12.5 percent of at-scale programs target teachers based on their experience. Other notable organizational characteristics include the provision of complementary materials such as storybooks and reading materials. Top-performing PD programs and at-scale PD programs are similar in the amount of materials they provide, but our results suggest that the kinds of complementary materials may differ somewhat. For instance, only 12.5 percent and 21 percent of at-scale programs provide storybooks and reading materials, respectively—materials correlated with student learning gains—as compared to 36 percent and 43 percent of evaluated programs.

Turning next to Content (table 5), top-performing PD programs and at-scale PD programs perform similarly. In both instances, the majority of programs include subject content and subject-specific pedagogy as either a primary or secondary focus. Few programs—none of the top performers—and only eight percent of at-scale programs lack a subject focus. Moreover, no top-performing programs and few at-scale programs (fewer than six percent) focus on general training in areas such as counseling or providing training on how to use a specific tool—types of training that are statistically linked to lower gains in student learning.

Finally, Delivery characteristics (table 5) include whether or not there are consecutive days of face-to-face training, training location, the amount of time teachers spend practicing with one another, and follow-up visits. Specifically, 100 percent of top-performing programs include consecutive days of face-to-face training as compared to 85 percent of evaluated programs. Our research further suggests that the location of PD training programs may influence program effectiveness, and

Table 5. Comparison of Means of At-Scale Programs and Top-Performing, Evaluated Programs

	Top performers	Obs	At-scale programs	Obs
Organization variables				
Robust characteristics				
Targeting by years of experience	13.33%	15	12.50%	48
Participation has implications for status, salary or promotion	87.50%	16	58.33%	48
Characteristics with large coefficients				
Program provides other reading materials (flashcards, word banks, reading pamphlets)	42.86%	14	20.83%	48
Program provides storybooks	35.71%	14	12.50%	48
Number of schools	148	13	6,367	29
Content variables				
Robust characteristics				
Focus is counseling	0%	16	3.60%	139
Focus is a specific tool	0%	16	6.47%	139
No subject focus	0%	16	8.33%	48
Training involves lesson enactment	62.50%	8	72.66%	139
Characteristics with large coefficients				
Focus is subject content	81.25%	16	27.34%	139
Subject focus is math	12.50%	16	54.17%	48
Subject focus is information technology	6.25%	16	22.92%	48
Delivery variables				
Robust characteristics				
Initial period of face-to-face training for several days in a row	100.00%	15	85.42%	48
Training held at central location including hotel conference room etc.	37.50%	16	72.97%	139
Includes distance learning	9.09%	11	NA	NA
Characteristics with large coefficients				
Proportion of face-to-face training spent practicing with teachers	39.81%	9	15.57%	34
Trainers are education university students	6.25%	16	0%	139
Follow-up visits to review material	12.50%	16	10.42%	48
Includes follow-up visits	84.62%	13	49.64%	139
Median Number of follow up visits	3.5	13	0	130

Source: Constructed by authors, comparing summary statistics for the top performing professional development (PD) programs among rigorously evaluated PD programs to at-scale PD programs.
Note: For the full list of statistics, see supplementary online appendix Tables A7.1–A7.3.

training held at central locations such as hotels or conference rooms (as opposed to universities or training centers) may be less effective. Currently 73 percent of at-scale, government-funded programs are held at central locations as compared to only 38 percent of evaluated programs.

Follow-up visits with teachers and the amount of time teachers spend practicing with other teachers during the training program are shown to be positively correlated with large coefficients (albeit not statistically significant) on student learning. In both instances, top-performing PD programs include more follow-up visits (five versus two median visits among programs with visits) and spend more time allowing teachers to practice with other teachers (40 percent versus 16 percent of training time) than do at-scale programs.[11] Results of our analysis suggest that training may be more effective if there are follow-up visits. This is an imperative finding when comparing top-performing PD programs, in which 85 percent include follow-up visits, with government-funded, at-scale PD programs, in which only half of programs include follow-up visits. Also, in top-performing PD programs, teachers spend more time practicing what they have learned with other teachers (40 percent of overall training time) relative to at-scale programs (only 16 percent). An existing body of research suggests that when teachers have opportunities to practice the new skills they acquire in PD programs, they are more likely to adopt these new skills in their classrooms (Wiley and Yoon 1995; Wenglinsky 2000; Angrist and Lavy 2001; Borko 2004).

Discussion

Governments spend enormous amounts of time and money on in-service professional development. Many countries have multiple in-service PD programs running simultaneously, as evidenced by our sample of at-scale PD programs. Many go unevaluated and may be ineffective. This paper makes three major contributions: first, it reveals broad weaknesses in reporting on teacher PD interventions. There are almost as many program types as there are programs, with variations in subject and pedagogical focus, hours spent, capacity of the trainers, and a host of other variables. Yet reporting on these often seeks to reduce them to a small handful of variables, and each scholar decides independently which variables are most relevant to report. We propose a standard set of indicators—the ITTSI—that would encourage consistency and thoroughness in reporting. Academic journals may continue to pressure authors to report limited information about the interventions, wishing instead to reserve space for statistical analysis. However, authors could easily include the full set of indicators in an appendix attached to the paper or online.

Second, this paper demonstrates that some characteristics of teacher PD programs—notably, linking participation to incentives such as promotion or salary implications, having a specific subject focus, incorporating lesson enactment in the training, and including initial face-to-face training—are positively associated with student test score gains. Furthermore, qualitative evidence suggests that follow-up visits to reinforce skills learned in training are important to effective training. Further documentation of detailed program characteristics, coupled with rigorous evaluation, will continue to inform effective evaluations.

The impacts of these characteristics are not small: having a specific subject focus and incorporating lesson enactment are associated with 0.24 and 0.10 more standard deviations in learning, respectively, for example. Comparing these effect sizes to those from a sample of 747 education-related randomized controlled trials in the United States puts them both above the 50th percentile in terms of effectiveness (Kraft 2020). Comparing to a set of 130 randomized controlled trials in low- and middle-income countries likewise put them at or above the 50th percentile of 0.10 standard deviations (Evans and Yuan 2020). In high-income countries, Kennedy (2019) proposes that the impact of teacher PD programs be benchmarked against a much less costly "community of practice" model in which teachers help each other, like Papay et al. (2020). While we are not aware of a rigorously evaluated, costed model of that class of program in a low- or middle-income country, an alternative would be to compare teacher PD results to a pure monitoring model, such as an increase in inspections. Along these lines, Muralidharan et al. (2017) show—using data from India—that increased frequency of monitoring would be a much more cost-effective way to reduce effective class sizes (through reduced teacher absenteeism) than hiring more teachers. These are useful avenues to pursue for future research as countries consider the cost-effectiveness of alternative investments in teachers.

Third, by comparing the means of at-scale PD programs with top-performing evaluated programs, our findings highlight gaps between what evidence suggests are effective characteristics of teacher PD programs and the contextual realities of most teacher PD programs in their design, content, and delivery. In particular, our findings taken together suggest that at-scale programs often lack key characteristics of top-performing training programs. At-scale programs are much less likely to be linked to career incentives, to provide storybooks or other reading materials, to have a subject content focus, to include time for practicing with other teachers, or to include follow-up visits.

The approach taken by this paper centers on using the ITTSI to collect and compare data on rigorously evaluated and at-scale, government-funded teacher PD programs. This approach has limitations. First, the evidence of what works within rigorously evaluated programs is limited by those programs that have been evaluated. There may be innovative PD programs that are not among the "top performers" simply because they have yet to be evaluated. While this evidence base can push policymakers away from approaches that do not work, it should not deter policymakers from innovating and evaluating those innovations.

A second, related limitation concerns the relatively small sample of evaluated teacher PD programs in low- and middle-income countries, on which our findings about effective PD characteristics are based. Some of the larger coefficients in the regressions are driven by a small number of teacher training programs. These instances have been noted in the text. As more evaluations of PD programs are conducted, the ITTSI can be applied to these and our analyses re-run to shed further light on the specific characteristics associated with PD programs that improve student learning. The ITTSI data were already updated once in this way in 2018, increasing the number of evaluated programs in our sample from 26 to 33.

Third, a conceptual concern with evaluating teacher professional programs is the risk that impacts may be explained by observer effects (also referred to as Hawthorne effects). These effects have been documented in education (Muralidharan et al. 2017) and health in low- and middle-income countries (Leonard 2008; Leonard and Masatu 2010). The impact of any education intervention may partly be due to observer effects, since the introduction of an intervention suggests that someone is paying attention to the teacher's efforts. Both randomized controlled trials and more traditional monitoring and evaluation may enhance these effects, as teachers may further respond favorably to the observation associated with measurement. Randomized controlled trials and quasi-experimental studies with a credible comparison group overcome part of this concern, as the observer effect associated with measurement will exist in both the treatment and comparison groups, and measured program impacts should be net of those effects.

That leaves the impact of the intervention itself. In this review, all of the studies we include evaluate interventions and, as such, all may be subject to an observer effect. Our analysis implicitly assumes the magnitude of this observer effect to be constant across different types of PD. By comparing PD characteristics across programs, we observe whether those characteristics are associated with a larger total effect on learning. Part of that total effect may stem from increased teacher skills, and part may be explained by certain PD characteristics inducing greater observer effects (since any observer effects that are uncorrelated with PD characteristics would be absorbed in our regression constant terms). In the short run, the impact for students is observationally equivalent. Even with longer run studies (of which there are very few in education and development), observer effects may fade, but teacher skills may also depreciate (Cilliers et al. 2020). As a result, we consider the total association of PD characteristics with student learning, including through increased teacher human capital and observer effects.

Fourth, there are challenges in comparing evaluated PD programs with at-scale PD programs. As the data demonstrate, at-scale PD programs tend to be larger programs designed by governments, often at the national level, and aimed at providing broad training to teachers. In light of these differences, we highlight the fact that top-performing programs—regardless of their core objectives—share certain common sets of characteristics that most at-scale programs do not share. Awareness of these characteristics may be useful in the conceptualization and implementation of future teacher PD programs in low- and middle-income countries, including large-scale programs funded by governments.

One key reason that at-scale programs may differ from successful, evaluated programs is that the latter group of evaluations may not be designed in a way that is conducive to scaling. Evaluated programs tend to be much smaller than at-scale programs: in our data, evaluated programs reached an average of 96 schools versus at-scale programs that reached more than 6,000 schools on average (supplementary online appendix table A7.1). These smaller programs often have higher per-pupil costs (Evans and Popova 2016b), so scaling them nationwide requires cutting elements. Smaller programs are easier to staff and easier to monitor. Evaluated programs were three times as likely to be designed by researchers and less than one-third as likely to be implemented by government (supplementary online appendix table A7.1). One solution, obviously, is more large-scale evaluations, like Loyalka et al. (2019). However, even smaller evaluations can do more to mimic scalable policies. Gove et al. (2017), reflecting on programs evaluated both at pilot and at scale in Kenya and Liberia, suggest the value of testing as many elements as possible in the pilot, using government systems in the pilot as much as possible, and to make sure that pilot costs are within what a government budget can handle. Duflo et al. (2020) combine these two approaches in a recent nationwide, five-arm randomized controlled trial in Ghana, to test the scalability of four different models to reach remedial learners, which had previously been tested in small pilot randomized controlled trials elsewhere. When implemented within existing government systems, they find all four interventions to be effective, pointing to the program's inception within the government as key, as opposed to an initial non-government organization initiative subsequently and imperfectly implemented by the government.

Improving in-service teacher professional development may be a clear win for governments. They are already spending resources on these programs, and there is broad support for these programs among teachers and teachers' unions. Interventions such as the above provide learning opportunities for country governments and stakeholders seeking to design effective teacher PD programs. While no single characteristic of top-performing PD programs may transform an ineffective PD program into an effective one, this paper highlights trends in top-performing programs, such as including incentives, a specific subject focus, and lesson enactment. These are characteristics that, if included and implemented successfully, have the potential to improve the quality of teacher PD programs, and ultimately, the quality of instruction and student learning.

Notes

Anna Popova is a doctoral candidate at Stanford University's Graduate School of Education (apopova@stanford.edu). David Evans is a senior fellow at the Center for Global Development (devans@cgdev.org). Mary Breeding (mbreeding@worldbank.org) and Violeta Arancibia (varancib@mac.com) are consultants with the World Bank.

The authors are grateful for comments from Denise Bello, Luis Benveniste, Barbara Bruns, Martin Carnoy, Joost de Laat, Margaret Dubeck, Deon Filmer, Susanna Loeb, Prashant Loyalka, Ezequiel Molina, Andrew Ragatz, and Halsey Rogers. They are also grateful to Fei Yuan for excellent research assistance, to Veronica Michel Gutierrez, Olga A. Rines, Lea Jeanne Marie Lungmann, Fata No, and Elissar Tatum Harati for their support with data collection, and to numerous teacher training implementers for providing information on programs. This paper subsumes an earlier paper, "Training Teachers on the Job: What Works and How to Measure It" (World Bank Policy Research Working Paper Number 7834).

This work was supported by the Bill & Melinda Gates Foundation, the World Bank's Systems Approach for Better Education (SABER) Trust Fund, which was supported by the United Kingdom's Department for International Development (DFID) and Australia's Department of Foreign Affairs and Trade (DFAT), and the Strategic Impact Evaluation Fund at the World Bank.

1. Both samples focus on teacher training programs at the primary and secondary school level. Preprimary schools are excluded.

2. The databases we searched were the Education Resources Information Center (ERIC); Academic Search Complete; Business Source Complete; Econlit with Full Text; Education Full Text (H. W. Wilson); Education Index Retrospective: 1929–1983; Education Source; Educational Administration Abstracts; Social Science Full Text (H. W. Wilson); Teacher Reference Center; and EconLit. We looked for articles containing the terms ("teacher training" OR "teacher education" OR "professional development") AND ("learning" OR "scores" OR "attainment") AND ("impact evaluation" OR "effects") AND ("developing country 1" OR "developing country 2" OR "developing country N"), where "developing country" was replaced by country names.

3. In six cases, program implementers failed to schedule an interview after three attempts at contact, and in the case of one older program, the implementer had passed away. Interviews were held over the phone or in-person, and lasted between 45 and 90 minutes for each program.

4. A limitation is that some of the impact estimates from school-randomized control trials in our evaluated sample are over-estimates because the authors fail to account for the clustering of children within teachers or schools (Hedges 2009).

5. For categorical variables, this is straightforward. For example, we convert the original categorical variable for the location of the initial teacher PD—which includes response options of schools, a central location, a training center, or online—into four dummy variables. In order to convert the continuous variables to a comparable scale, we create a dummy for each continuous variable which, for a given program, takes a value of 1 if the continuous variable is greater than the median value of this variable across all programs, and a value of 0 if it is less than or equal to the value of this variable across all programs. We apply this method to the conversion of all continuous variables except three—proportion of teachers that dropped out of the program, number of follow-up visits, and weeks of distance learning—which we convert directly to dummy variables that take a value of 1 if the original variable was greater than 0, and a value of 0 otherwise.

6. We estimate bootstrapped standard errors by resampling our data with replacement 1,000 times. We run Fisher randomization tests by treating each indicator PD characteristic as a treatment and calculating a randomization distribution of mean differences (the test statistic) across treatment assignments. Specifically, for 1,000 permutations, we randomly reassign values of 0 or 1 to the independent variables in our regressions, while maintaining the overall proportion of 0s and 1s observed in the empirical sample for a given variable. We then calculate Fisher exact *p*-values by finding the proportion of

the randomization distribution that is larger than our observed test statistic (Fisher 1925, 1935; Imbens and Rubin 2015).

7. These regions include: Africa, Eastern and Central Europe, Latin American and the Caribbean, the Middle East and North Africa, and East and South Asia.

8. This includes programs ongoing in 2016 and programs that were implemented anytime in the range of 2012 to 2016. Hence, the programs could have been designed prior to 2012. We still include them if they were implemented any time between 2012 and 2016. We were not successful in obtaining roster information in all countries. For instance, in Morocco and the Arab Republic of Egypt, the Ministries of Education were in the process of making changes to the structure and delivery of teacher training programs and indicated that it was not a good time for data collection. In Tanzania there was a change in leadership among government counterparts during efforts to complete the roster and data collection process, and we were not able to properly sample and apply the ITTSI in all teacher-training programs in the country. In India, we had initially identified two states, Bihar and Karnataka, to work with at the subnational level, but ultimately only collected data in one state, Bihar, since the principal government counterpart in Karnataka was not available to complete the roster.

9. In some cases, we test a negative (e.g., no implications for status in table 2 or no subject focus in table 3) because we are testing an exhaustive series of indicators derived from the same question (e.g., subject focus is math, subject focus is literacy, or no subject focus).

10. Because it is difficult to imagine an effective teacher professional development program that teachers actively dislike (they have to learn for it to work, after all), their preferences are relevant.

11. When we include programs with no follow-up visits, the median number of follow-up visits to teachers in top programs becomes 3.5 as compared to 0 for at-scale programs.

References

Albornoz, F., M. V. Anauati, M. Furman, M. Luzuriaga, M. E. Podestá, and I. Taylor. 2018.“Training to Teach Science: Experimental Evidence from Argentina.” Policy Research Working Paper 8594, World Bank, Washington, DC.

Angrist, J. D., and V. Lavy. 2001. “Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools.” *Journal of Labor Economics* 19 (2): 343–69.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady. 2016. “Teacher Quality and Learning Outcomes in Kindergarten.” *Quarterly Journal of Economics* 131 (3): 1415–53.

Baker, S., and S. Smith. 1999. “Starting off on the Right Foot: The Influence of Four Principles of Professional Development in Improving Literacy Instruction in Two Kindergarten Programs.” *Learning Disabilities Research & Practice* 14 (4): 239–53.

Banerjee, A., S. Cole, E. Duflo, and L. L. Linden. 2007. “Remedying Education: Evidence from Two Randomized Experiments in India.” *Quarterly Journal of Economics* 122 (3): 1235–64.

Berlinski, S., and M. Busso. 2017. “Challenges in Educational Reform: An Experiment on Active Learning in Mathematics.” *Economics Letters* 156: 172–5.

Bold, T., D. Filmer, G. Martin, E. Molina, C. Rockmore, B. Stacy, J. Svensson, and W. Wane. 2017. “What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa.” Policy Research Working Paper 7956, World Bank, Washington, DC.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. *Introduction to Meta-Analysis.* Chichester, United Kingdom: John Wiley & Sons.

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

Bourgeois, E., and J. Nizet. 1997. *Aprendizaje y formación de personas adultas.* Paris, France: Presses Universite de France.

Cardemil, C. 2001. "Procesos y condiciones en el aprendizaje de adultos." *Jornada Nacional de Supervisores. Supervisión para aprendizajes de calidad y oportunidades para todos. Educación Rural.* Santiago: Ministerio de Educación. https://repositorio.uahurtado.cl/handle/11242/8517.

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

Cilliers, J., B. Fleisch, J. Kotze, M. Mohohlwane, and S. Taylor. 2020. " The Challenge of Sustaining Effective Teaching: Spillovers, Fade-out, and the Cost-effectiveness of Teacher Development Programs." Unpublished Working Paper. https://www.dropbox.com/s/6xmv7283oxoysj2/The%20Challenge%20of%20Sustaining%20Effective%20Teaching%20with%20appendix.pdf?dl=0.

Darling-Hammond, L., R. C. Wei, A. Andree, N. Richardson, and S. Orphanos. 2009. *Professional Learning in the Learning Profession.* Washington, DC: National Staff Development Council. https://edpolicy.stanford.edu/sites/default/files/publications/professional-learning-learning-profession-status-report-teacher-development-us-and-abroad.pdf.

Desimone, L. M. 2009. "Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures." *Educational Researcher* 38 (3): 181–99.

Duflo, A., J. Kiessel, and A. Lucas. 2020. "External Validity: Four Models of Improving Student Achievement." Working Paper No. w27298, National Bureau of Economic Research, Cambridge, MA.

Evans, D. K., and A. Popova. 2016a. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31 (3): 242–70.

———. 2016b. "Cost-Effectiveness Analysis in Development: Accounting for Local Costs and Noisy Impacts." *World Development* 77: 262–76.

Evans, D. K., and F. Yuan. 2020. "How Big are Effect Sizes in International Education Studies?" Working Paper 545, Center for Global Development, Washington, DC.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*, first edition. Edinburgh: Oliver and Boyd Ltd.

——— 1935. *The Design of Experiments*, sixth edition. Edinburgh: Oliver and Boyd, Ltd, 1951.

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

Gersten, R., M. J. Taylor, T. D. Keys, E. Rolfhus, and R. Newman-Gonchar. 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. http://files.eric.ed.gov/fulltext/ED544681.pdf.

Gove, A., M. K. Poole, and B. Piper. 2017. "Designing for Scale: Reflections on Rolling Out Reading Improvement in Kenya and Liberia." *New Directions for Child and Adolescent Development* 2017 (155): 77–95.

Hedges, L.V. 2009. "Effect Sizes in Nested Designs." In *The Handbook of Research Synthesis and Meta-analysis*, edited by H. Cooper, L. V. Hedges and J. C. Valentine, 337–56. New York, NY: Russell Sage Foundation.

Huberman, M. 1989. "The Professional Life Cycle of Teachers." *Teachers College Record* 91 (1): 31–57.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge, UK: Cambridge University Press.

Kennedy, M. M. 2019. "How We Learn About Teacher Learning." *Review of Research in Education* 43 (1): 138–62.

Kerwin, J. T., and R. L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *Review of Economics and Statistics* 103 (2): 251–64.

Knowles, M. S., E. F. Holton, and R. A. Swanson. 2005. *The Adult Learner*, sixth edition. Burlington, MA: Elsevier.

Kraft, M. A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher* 49 (4): 241–53.

Kraft, M. A., and J. P. Papay. 2014. "Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience." *Educational Evaluation and Policy Analysis* 36 (4): 476–500.

La Paro, K. M., and R. C. Pianta. 2003. *CLASS: Classroom Assessment Scoring System*. Charlottesville, VA: University of Virginia.

Leonard, K. L. 2008. "Is Patient Satisfaction Sensitive to Changes in the Quality of Care? An Exploitation of the Hawthorne Effect." *Journal of Health Economics* 27 (2): 444–59.

Leonard, K. L., and M. C. Masatu. 2010. "Using the Hawthorne Effect to Examine the Gap Between a Doctor's Best Possible Practice and Actual Performance." *Journal of Development Economics* 93 (2): 226–34.

Loyalka, P., A. Popova, G. Li, and Z. Shi. 2019. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program." *American Economic Journal: Applied Economics* 11 (3): 128–54.

McEwan, P. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353–94.

Molina, E., S. F. Fatima, A. Ho, C. M. Hurtado, T. Wilichowski, and A. Pushparatnam. 2018. "Measuring Teaching Practices at Scale: Results from the Development and Validation of the Teach Classroom Observation Tool." Policy Research Working Paper 8653, World Bank, Washington, DC.

Muralidharan, K., J. Das, A. Holla, and A. Mohpal. 2017. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." *Journal of Public Economics* 145: 116–35.

Papay, J. P., E. S. Taylor, J. H. Tyler, and M. E. Laski. 2020. "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data." *American Economic Journal: Economic Policy* 12 (1): 359–88.

Piper, B., and M. Korda. 2011. *EGRA Plus: Liberia (Program evaluation report)*. Durham, NC: RTI International.

Rogers, H., and A. Demas. 2013. *The What, Why, and How of the Systems Approach for Better Education Results (SABER)*. Washington, DC: World Bank. http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/Background/SABER_Overview_Paper.pdf.

Shulman, L. S. 1986. "Those Who Understand: Knowledge Growth in Teaching." *Educational Researcher* 15 (2): 4–14.

TNTP. 2015. *The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development*. The New Teacher Project. http://files.eric.ed.gov/fulltext/ED558206.pdf.

Villegas-Reimers, E. 2003. *Teacher Professional Development: An International Review of the Literature*. Paris: UNESCO International Institute for Educational Planning. http://www.iiep.unesco.org/en/publication/teacher-professional-development-international-review-literature.

Walter, C., and J. Briggs. 2012. *What Professional Development Makes the Most Difference to Teachers*. Oxford: University of Oxford Department of Education. https://www.oupjapan.co.jp/sites/default/files/contents/events/od2018/media/od18_Walter_reference.pdf.

Wenglinsky, H. 2000. "How Teaching Matters: Bringing the Classroom Back into Discussions of Teacher Quality." Policy Information Center Report, Educational Testing Service (ETS).

Wiley, D., and B. Yoon. 1995. “Teacher Reports of Opportunity to Learn: Analyses of the 1993 California Learning Assessment System.” *Educational Evaluation and Policy Analysis* 17 (3): 355–70.

Wood, F. H., and F. McQuarrie, Jr. 1999.“On the Job Learning. New Approaches will Shape Professional Learning in the 21st Century.” *Journal of Staff Development* 20: 10–13.

Yoon, K. S., T. Duncan, S. W. Y. Lee, B. Scarloss, and K. Shapley. 2007. *Reviewing the Evidence on how Teacher Professional Development Affects Student Achievement* (Issues & Answers Report No. 033). 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. http://files.eric.ed.gov/fulltext/ED498548.pdf.