Policy Research Working Paper 8572

Teacher Professional Development around the World

The Gap between Evidence and Practice

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

WORLD BANK GROUP
Africa Region
Office of the Chief Economist
August 2018

Policy Research Working Paper 8572

Abstract

Teachers, like all professionals, require ongoing professional development opportunities to improve their skills. This paper provides evidence on effective professional development characteristics and how at-scale programs incorporate those characteristics. The authors propose a standard set of 70 indicators—the In-Service Teacher Training Survey Instrument—for reporting on professional development programs as a prerequisite for understanding the characteristics of those programs that improve student learning. The authors apply the instrument to rigorously evaluated professional development programs in low- and middle-income countries. Across 33 programs, those 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. The authors then apply the instruments to a sample of 139 government-funded, at-scale professional development programs across 14 countries. This analysis uncovers a sharp gap between the characteristics of teacher professional development programs that evidence suggests are effective and the global realities of most teacher professional development programs.

This paper is a product of the Office of the Chief Economist, Africa Region and the World Development Report 2018 team. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at apopova@stanford.edu and devans2@worldbank.org.

Produced by the Research Support Team

Teacher Professional Development around the World:

The Gap between Evidence and Practice

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

JEL Codes: I20, J24, O10

Keywords: Education quality, teacher training, professional development, economic development

Acknowledgments: The authors are grateful for comments from Denise Bello, Luis Benveniste, Tara Beteille, Barbara Bruns, Martin Carnoy, Joost de Laat, Margaret Dubeck, Deon Filmer, Susanna Loeb, Prashant Loyalka, Ezequiel Molina, Andrew Ragatz, and Halsey Rogers, and for financial support from the Bill and Melinda Gates Foundation and the Strategic Impact Evaluation Fund at the World Bank. 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). In 2017 and 2018, the paper underwent a major update, adding new evaluations of teacher professional development programs and including new data on in-service teacher training programs from a sample of education systems around the world.

1. Introduction

Students around the world – especially in low- and middle-income countries – are not learning enough in school. In a recent international assessment, students at the 75th percentile of performance in selected countries from the Caribbean, North Africa, and Eastern Europe performed below the 25th percentile average for high-income countries (World Bank, 2018). A growing body of research demonstrates that teachers are among the most important determinants of student learning. The difference between a weak teacher and an excellent teacher has been measured at up to a full year of student learning in the United States (Hanushek and Rivkin, 2010), and large teacher effects have also been documented in Chile (MINEDUC, 2009), Ecuador (Araujo et al., 2016), Pakistan (Talance, 2015), and Uganda (Buhl-Wiggers et al., 2017). Beyond immediate improvements in student learning, teachers who raise student test scores significantly improve students' long-term outcomes – such as their probability of graduating college and adult salaries – and decrease the likelihood of teenage pregnancy (Chetty, Friedman, & Rockoff, 2014).

Furthermore, in recent reviews of the education literature, improving pedagogy so that it is more directed to individual student levels – an action that depends significantly on teachers either carrying out formative assessments or targeting instruction – and providing individualized, repeated teacher training, associated with a specific method or task were among the most recommended interventions for improving student learning (Evans and Popova, 2016; Kremer, Brannen, & Glennerster, 2013). For example, the full Early Grade Reading Assessment (EGRA) program in Liberia, which trained teachers to use an initial reading assessment and then continually assess student performance, increased students' reading comprehension dramatically (Piper & Korda, 2011).

In-service teacher professional development is important to evaluate even beyond the promising evidence from a collection of evaluations, which show that it can – when designed correctly – improve student learning. Teacher professional development can take many forms ranging from traditional, government-mandated mass training programs to teacher pedagogical support groups headed by coaches or mentors that provide needs-based, embedded support. Significant government and donor resources are funneled into training programs. Of 171 World Bank projects with education components between 2000 and 2012, nearly two-thirds included professional development to support teachers. Despite the significant resources spent on in-service teacher PD programs, rigorous evidence on the effectiveness of such programs remains limited. Overall, evidence for the small share of programs that have been evaluated is mixed, and it is often reported that most current teacher education programs are outdated and over-theoretical.

At the same time, many evaluations fail to provide sufficient details on the actual content or delivery mechanisms of the trainings to inform the design of successful programs. This may be due – in part – to a lack of instruments designed to measure teacher professional development. Instruments exist to capture the design of teacher policy on one end of the policy-practice spectrum,[1] and how teachers behave in the classroom at the other end of the spectrum.[2] However, to date, there exists no instrument to capture the step between teacher policy design and teachers' classroom practice; that is, how teachers are actually trained and which specific components of this training effectively improve teacher behavior and subsequently student learning. This is true despite the fact that the evidence we have suggests that there is much more variation in effectiveness across teacher training programs than across education programs more broadly (Evans and Popova, 2016; McEwan, 2015). In other words, teacher training programs vary

[1] See SABER Teachers (World Bank, 2013).

[2] Among others, see the Classroom Assessment Scoring System or CLASS (La Paro & Pianta, 2003), the Danielson Framework for Teaching Evaluation Instrument (Danielson, 2013), the Protocol for Language Arts Teaching Observation or PLATO (Grossman et al., 2013), and Stallings (Stallings, 1977).

enormously, both in their form and in their effectiveness.

This paper has three objectives. The first is to fill the information gap on the essential characteristics of PD programs by proposing a survey instrument – the In-Service Teacher Training Survey Instrument (ITTSI) – to document the design and implementation details of in-service teacher training programs. We piloted the instrument on a sample of PD programs from low- and middle-income countries whose impact has already been evaluated, and we analyzed the resulting data using a combination of quantitative and qualitative methods. The second objective is to use those data to characterize the current evidence on in-service teacher training in low- and middle-income countries, comparing the characteristics of programs that resulted in large student learning gains with those that did not. The third objective is to assess current in-service teacher PD practices from a sample of at-scale teacher PD programs across the world, and to analyze how current practice diverges from the best practices coming out of evaluated programs. After piloting the ITTSI instrument on evaluated PD programs and refining it, we used the resulting instrument to collect data from a sample of 139 recent PD programs across 14 low- and middle-income countries.

Findings from applying the ITTSI to evaluated PD programs suggest that characteristics positively associated with program impact on student learning include 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, among others. Meanwhile, program implementers themselves most commonly mention the provision of mentoring follow-up visits, engaging teachers for their opinions and ideas, and designing programs in response to local context as being responsible for positive impacts on student learning.

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 training supported by research evidence and that currently being provided by many government-funded PD programs.

This paper proceeds as follows. Section 2 summarizes a sample of the theoretical literature on in-service teacher training and provides insights from impact evaluations of programs in high-income countries. Section 3 describes the methodological approach, including the instrument design, the search strategy for evaluated programs, the sampling strategy for at-scale unevaluated programs, the data collection for each of these samples, and the analytical strategy. Section 4 presents the results of our quantitative and qualitative analyses, and Section 5 concludes.

2. Background

Theory

The education literature suggests at least five factors to consider in the design of teacher PD programs: the students; the teachers; and the method, content, and duration of instruction. First, because working, professional teachers are the students in PD, principles of adult education are relevant. Adult education tends to work best with clear applications rather than a theoretical focus (Cardemil, 2001; Knowles, Holton, & Swanson, 2005). This relates both to how practical the delivery of PD is, and to targeting, a seemingly important overarching aspect of PD programs. 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, so there may be a benefit from leveraging that time (TNTP, 2015).

Second, the quality of instructors – i.e., those providing the PD – is crucial to learning (Knowles, Holton, & 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, the method of instruction should include concrete, realistic goals (Baker & Smith, 1999) and the teaching of formative evaluation so that teachers can effectively evaluate their own progress towards their teaching goals (Bourgeois & Nizet, 1997). Fourth, 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 against brief, one-time workshops and in favor of sustained contact over a significant period of time (Desimone, 2009).

Fifth, on the content of PD – relative to theory or general pedagogy, subject-specific pedagogy is likely to be most effective, as different subjects require radically different pedagogies (Villegas-Reimers, 2003). For example, a more scripted approach may work for early grade reading, whereas later grade science will require higher-order thinking skills. Sixth, on the location of instruction, another important delivery characteristic, teacher PD in the school ("embedded") is likely to be most effective so that concrete problems faced in the local environment can be raised, and teachers can receive feedback on actual teaching (Wood & 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 & Papay, 2014).

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. A recent meta-analysis of 196 randomized field experiments to improve education in the U.S. – that measure student test scores as an outcome – examined the impact of both "general" and "managed" professional development, relative to a wide range of other interventions such as tutoring and teacher incentives (Fryer, 2017). General professional development (PD) – as the name suggests – leaves a fair amount of flexibility, even as it may focus on classroom management or increasing the rigor of teachers' knowledge. Managed professional development, on the other hand, is much more prescriptive; it 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. However, the analysis is based on relatively few studies, with just seven general PD studies and two managed PD studies. Nonetheless, this finding in support of specific and practical teacher training aligns with that of Walter & Briggs (2012) who, in a review of 35 evidence-based studies of teacher PD, found that concrete and classroom-based programs make the most difference to teachers.

Another U.S.-focused review 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). But this review also draws on few strong studies: of the 1,300 studies included in this review, only nine had pre- and post-test data and some sort of control group. Similarly, a 2014 review of professional development 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. The authors concluded, "The limited research on effectiveness means that schools and districts cannot use evidence of effectiveness alone to narrow their choice" (Gersten et al., 2014). As such, we look also to a wider range of evidence. For example, one recent US review which includes qualitative as well as quantitative studies concludes that teacher training is most effective when it focuses on "concrete tasks of teaching, assessment, observation and reflection" instead of abstract teaching concepts. Effective programs were not "one-shot workshops" – but rather embedded in the curriculum; and they were sustained and intense (Darling-Hammond et al., 2009).

Much of the evidence from other high-income countries is qualitative. Narrative empirical analysis by Darling-Hammond, Wei, & Andree (2010) highlighted that in high-achieving countries, in-service support to teachers includes (a) mentoring for all beginners, coupled with a reduced teaching load and shared planning time for new and mentor teachers; (b) extensive opportunities for ongoing professional learning, embedded in substantial planning and collaboration time at school; and (c) teacher involvement in curriculum and assessment development and decision making. For example, teaching in Japan includes a practicum year for all beginning teachers during which teachers have a reduced teaching load, attend in-school training with guidance teachers twice a week, and receive weekly out-of-school training, including seminars and visits to other schools (Darling-Hammond, 2005). The Japanese education system also includes a lesson study approach to professional development, in which teachers rotate in preparing and teaching lessons addressing a specific goal of their choosing, while others observe and record the lesson, and subsequently provide feedback and make suggestions for improvement (Darling-Hammond et al., 2010). Singapore's Teachers Network learning circles – in which between four and ten teachers and a facilitator meet for eight two-hour sessions over a period of four to twelve months, to collaboratively identify and solve common problems using discussions and action research – similarly encourage teachers to be reflective practitioners (Darling-Hammond et al., 2010).

Limited high-quality experimental or quasi-experimental evidence makes it difficult to draw detailed conclusions about what works within teacher training even in rich countries. However, from a combination of the more rigorous quantitative and qualitative studies above, there is suggestive evidence that in-service teacher PD programs in high-income countries have been most effective at improving student learning where they have been embedded in the curriculum; prescribed a specific method, with detailed instructions on implementation; included significant and sustained in-person follow-up support for teachers; and involved teachers in a co-learning model.

3. Method

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 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 and student assessment results—and one sample of *at-scale*, government-funded PD programs. Below we discuss the design of the survey instrument and both data sets. The remainder of this section describes the methodology for each of the three aforementioned research phases.

Designing the In-Service Teacher Training Survey Instrument (ITTSI)

The ITTSI was designed based on (a) the descriptive, impact evaluation, and theoretical literatures characterized in the previous section, and (b) the authors' prior experience studying in-service teacher training. Drawing on these, we drafted a list of key indicators to capture details about a range of program characteristics falling into four categories: Overarching Aspects, Content, Delivery, and Perceptions (Figure 1).

Taking each of these in turn, the Overarching Aspects 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 materials it provides, the scale of the program, and its cost. Content includes indicators capturing the type of knowledge or skills that a given program aims to build among beneficiary teachers, for example, whether the program focuses on subject content (and if so, which subject), pedagogy, new technology, classroom management, counseling, assessment, or some combination of these.

Delivery focuses on indicators capturing program implementation details, such as whether it is delivered through a cascade model (where the program trains trainers who in turn train the teachers, sometimes with additional layers in between), the profile of the trainers who directly train the teachers, the location of the training, the size of 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.

During the first phase of data collection, in which we coded the information reported in our sample of impact evaluations, as we learned more about the programs we added new indicators to our instrument and adjusted existing ones in an iterative process so as to accurately characterize the full range of programs. This resulted in a draft instrument consisting of a total of 51 indicators. We piloted this draft instrument by using it to collect data on a sample of evaluated programs and analyzed these data to see which program characteristics are most predictive of student learning. The results of this analysis are reported in this paper. To validate the ability of the ITTSI to capture a detailed picture of PD programs, subsequent to this data collection and analysis, we shared our results with a series of expert researchers and practitioners in teacher PD and updated the indicators based on their feedback, including the addition of a series of questions specific to online programs. The resulting final version of the instrument, which includes 70 indicators plus three pieces of meta-data, is presented in Appendix A. This version of the ITTSI was translated into four different languages—Russian, French, Spanish, and Arabic—and subsequently used to collect data for the sample of at-scale PD programs. As a result, the number of indicators used to characterize at-scale programs (70) is different from the number of indicators used to characterize rigorously evaluated programs (51). However, all 51 original indicators are included among the 70 and allow for comparison across the two samples.

In addition to the full ITTSI, we designed a Brief In-Service Teacher Training Instrument (BITTSI). The BITTSI comprises a subset of questions from the ITTSI which the authors deemed most critical to ask of teacher training program coordinators, given the available research evidence and analysis of evaluated PD programs. The BITTSI covers 27 indicators – all also covered in the

ITTSI – and was utilized to collect data for the full sample of at-scale PD programs, while the ITTSI was applied to the three or four largest PD programs per country.

Applying the ITTSI to Evaluated PD Programs

Search Strategy

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 in terms of the impact they have on student learning. The resulting sample would serve first to inform a review of what we know to date about the effectiveness of different kinds of in-service teacher PD in those settings. Secondly, we used this sample of studies to inform the design of our survey instrument – by including questions about relevant characteristics either reported or noticeably omitted by studies – and to pilot the instrument in interviews with the program implementers.

Our inclusion criteria for the search were impact evaluations – i.e., studies that sought to establish the causal impact of a program through use of a counterfactual, including experimental and quasi-experimental studies – of primary education interventions in low- and middle-income countries that (a) focused primarily on in-service teacher training 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 do not explicitly restrict by year of authorship.

In order to identify papers fulfilling the above criteria, we searched 10 meta-databases through EBSCOhost: 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.

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, 2016). 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 us 42 full articles to assess their eligibility. After going through the full texts, another 18 papers were excluded as 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 33 programs. The search process is detailed in Figure 2. The 30 papers are listed in Appendix B.

Data Collection

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 provide. The draft version of the instrument for which we collected data included 51 indicators in total, and on average, information on 26 (51%) 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 overarching aspects 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% and 42% of information concerning program content and delivery, respectively, is reported. This is of particular concern given that theory suggests certain aspects of delivery are crucial, such as how much practice the program involves and whether or not it is delivered through a cascade model.

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 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.

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%) 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% for Overarching Aspects indicators,

[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.

68% of Content indicators, and 72% of Delivery indicators (Table 1).

Analytical Strategy

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:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where Y is the standardized impact on student test scores, X is an indicator of a given program characteristic, α is the constant, β is the coefficient on X, ε is the estimation error, and i represents each in-service teacher PD program in the sample. 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.

In preparation for this analysis, we standardize the impact estimates, Y, for each of the programs. We convert the independent program variables, X, 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. We standardize these effects and the associated standard errors in order to be able to compare them directly.[4]

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?). In order 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

[4] Our unit of analysis for effect size is an experimental or quasi-experimental pair, where a group of students taught by teachers who participated in a given PD program is compared to a control group taught by teachers who did not participate. Almost all the studies in our sample used difference-in-differences methods to estimate the effect of the teacher PD programs – or the larger programs of which training is a sub-component – on student learning and reported the effect size as a raw mean difference, *D*, between treatment and control groups, before and after a given program. Following Borenstein et al. (2009), we calculate the standardized effect size or mean difference, *d*, for each estimate, by dividing the raw mean difference, *D*, by the pooled standard deviation, S_{pooled}, as follows:

$$d = \frac{D}{S_{pooled}} \qquad \text{(Equation 1)}$$

S_{pooled} is the within-estimate standard deviation for the treatment and control groups combined. Where this is not directly reported in the studies we calculate it using the following equation derived from Borenstein et al. (2009):

$$S_{pooled} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, SE_D \qquad \text{(Equation 2)}$$

where n_1 is the sample size for the treatment group, n_2 is the sample size for the control group, and SE_D is the standard error of the raw mean difference. For a complete derivation of Equation 2 please see the mathematical appendix in Appendix C.

[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.

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] 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.

Applying the ITTSI to At-Scale PD Programs

Sample

The sampling process for at-scale programs is detailed in Figure 3. To obtain a sample of at-scale, government funded PD programs across the world we first identified four to five countries

[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 & Rubin, 2015).

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. 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 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 schools teachers). In some countries, such as Mexico and India where policy making about teacher professional development 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

[7] These regions include: Africa, Eastern and Central Europe, Latin American and the Caribbean, 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 the Kingdom of Morocco and the Arab Republic of Egypt, the Ministries of Education are 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 it was not possible 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.

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.

Data Collection

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%) 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%) indicators for the 91 programs to which the BITTSI was applied.

Analytical Strategy

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.

4. Results

This section characterizes the specific characteristics of teacher PD programs that successfully improve student learning in low- and middle-income countries, and it examines 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?

Quantitative Analysis

We discuss, for each of our categories – Overarching Aspects, Content, and Delivery – those characteristics we observe to be most associated with student learning gains, both characteristics that are robust according to our empirical strategy as well as characteristics with substantially large coefficients, as suggestive evidence. Tables 2, 3, and 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 Overarching Aspects (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. In the evaluated sample, in teacher PD programs where participation has no implications for promotion, salary, or status increases, student learning is 0.12 standard deviation lower (significant at 95%). Targeting participant teachers by their years of experience has the next largest, robust association with student learning, at 0.10 standard deviation higher (significant at 90%). 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, Cole, Duflo, & Linden, 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 newer, less-experienced teachers (Albornoz et al. 2017). 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. 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 deviation), 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 deviation lower impact on student learning (significant at 99%). 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 deviation in student learning (significant at 99%). 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%).

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%). 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%). Third, teacher PD trainings that are held remotely using distance learning are associated with a 0.10 standard deviation decrease in student learning (significant at 90%). 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 standard deviation 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.

Qualitative Analysis

We supplement the quantitative results with an analysis of self-reported perceptions by the implementers of the evaluated programs about 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 – 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 as well as 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. 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.

How Do At-Scale PD Programs Compare to Evaluated Top-Performers?

Government-funded, at-scale teacher PD programs differ sharply from programs that are evaluated in general, as well as from top-performing evaluated programs specifically. Evaluated programs are more likely than at-scale programs to focus on subject content (64% vs. 27%) or on pedagogy (58% vs. 37%). In both sets of programs, between 50 and 60 percent are delivered via a cascade training model. However, on average, evaluated PD programs – relative to at-scale programs – are more commonly targeted by grade (81% vs. 31%), linked to some sort of career incentives (42% vs. 17%), and are much longer in duration (60 hours vs. 13 hours). We provide a full list of average characteristics of at-scale programs and all evaluated programs (not just top-performers) in Appendix Tables D1-D3.

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 substantive and robust standard deviation increases in student assessment scores. In Tables 5-7, 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 Overarching Aspects (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% of top-performing PD programs link training to status or to new career opportunities such as promotion or salary, as compared to only 55% of at-scale programs. Our results suggest that without incentives, trainings 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% of top-performers and 12.5% of at-scale programs target teachers based on their experience. Other notable overarching 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% and 21% of at-scale programs provide storybooks and reading materials, respectively—materials correlated with student learning gains—as compared to 36% and 43% of evaluated programs.

Turning next to Content (Table 6), 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 8% of at-scale programs lack a subject focus. Moreover, no top-performing programs and few at-scale programs (fewer than 6%) focus on general trainings 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, the last set of characteristics – Delivery characteristics (Table 7) – 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% of top-performing programs include consecutive days of face-to-face training as compared to 85% of evaluated programs. Our research further suggests that the location of PD training programs may influence program effectiveness, and trainings held at central locations such as hotels or conference rooms (as opposed to universities or training centers) may be less effective. Currently 73% of at-scale, government-funded programs are held at central locations as compared to only 38% of top-performing programs.

The amount of time teachers spend practicing with other teachers during the training program and follow-up visits with teachers are shown to be positively correlated with large coefficients (albeit not statistically significant) with gains in student learning. In both instances, top-performing PD programs include more follow-up visits (5 versus 2 visits) and spend more time allowing teachers to practice with other teachers (40% versus 16% of training time) than do at-scale programs. 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% include follow-up visits with government-funded, at-scale PD programs, in which only half of programs include follow-up visits. Among programs that have follow-up visits, the median number of visits to teachers in top programs is five as compared to two for at-scale programs.[9] Also, in top-performing PD programs, teachers spend more time practicing what they have learned

[9] 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.

with other teachers (40% of overall training time) relative to at-scale programs (only 16%). 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 (Angrist & Lavy, 2001; Borko, 2004; Cohen & Hill ,1997; Wenglinsky, 2000; Wiley & Yoon, 1995; World Bank, 2012).

5. 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 the 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.

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 professional development programs that are not among the “top performers” simply because they have yet to be evaluated. While this evidence base can push policy makers away from approaches that do not work, it should not deter policy makers 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 rerun 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, 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 trainings 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. 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.

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.

References

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics, 19*(2), 343-369.

Baker, S., & Smith, S. (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 and Practice*, *14*(4), 239-253. Doi:10.1207/sldrp1404_5

Banerjee, A., Cole, S., Duflo, E., & Linden, L. L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics, 122* (3), 1235-1264. Doi:10.1162/qjec.122.3.1235

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (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., & Nizet, J. (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. Retrieved from http://biblioteca.uahurtado.cl/ujah/Reduc/pdf/pdf/mfn253.pdf. Accessed December 16th, 2016.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633-2679. Doi:10.1257/aer.104.9.2633

Danielson, C. (2013). *The Framework for Teaching: Evaluation Instrument.* The Danielson Group.

Darling-Hammond, L. (2005). Teaching as a profession: Lessons in teacher preparation and professional development. *Phi delta kappan*, *87*(3), 237. Doi:10.1177/003172170508700318

Darling-Hammond, L. Wei, R. C. & Andree, A. (2010). *How high-achieving countries develop great* *teachers*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from http://www.oup.hu/howhigh_doug.pdf. Accessed December 16th, 2016.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession.* Washington, DC: National Staff Development Council. Retrieved from http://www.ostrc.org/docs/document_library/ppd/Professionalism/Professional%20Learning%20in%20the%20Learning%20Profession.pdf. Accessed December 16th, 2016.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199.

Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *World Bank Research Observer*. Doi:10.1093/wbro/lkw004

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Edinburgh: Oliver and Boyd Ltd.

Fisher, R. A. (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. In *Handbook of Economic Field Experiments, 2*: 95-322.

Gersten, R., Taylor, M. J., Keys, T.D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches.* Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, U.S. Department of Education; and Regional Educational Laboratory Southeast at Florida State University. Retrieved from http://files.eric.ed.gov/fulltext/ED544681.pdf. Accessed December 16th, 2016.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445-470.

Hanushek, E. A., & Rivkin, S.G. (2010). *Using value-added measures of teacher quality* (CALDER Brief

No. 9). National Center forAnalysis of Longitudinal Data in Education Research. Retrieved from http://files.eric.ed.gov/fulltext/ED509683.pdf. Accessed December 16th, 2016.

Huberman, M. (1989). The professional life cycle of teachers. *Teachers College Record*, *91*(1), 31-57.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Knowles, M. S., Holton, E. F., & Swanson, R. A. (2005). *The adult learner* (6th ed.). Burlington, MA: Elsevier.

Kraft, M. A., & Papay, J. P. (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. Doi:10.3102/0162373713519496

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*, 297-300. Doi:10.1126/science.1235350

La Paro, K. M., & Pianta, R. C. (2003). *CLASS: Classroom assessment scoring system*. Charlottesville, VA: University of Virginia.

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

MINEDUC. (2009). *Resultados nacionales SIMCE 2008*. Santiago: Ministerio de Educación.

Piper, B., & Korda, M. (2011). *EGRA Plus: Liberia* (Program evaluation report). Durham, NC: RTI International. Doi:10.1016/0742-051X(89)90027-9

Stallings, J. (1977). *Learning to look: A handbook on classroom observation and teaching models*. Belmont, CA: Wadsworth Publishing Company.

TNTP. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. The New Teacher Project. Retrieved from http://files.eric.ed.gov/fulltext/ED558206.pdf. Accessed December 16th, 2016.

Villegas-Reimers, E. (2003). *Teacher professional development: An international review of the literature.*

Paris: UNESCO International Institute for Educational Planning. Retrieved from http://xa.yimg.com/kq/groups/23312027/1804288575/name/teacher+professional+development+an+international+review+from+the+literature.pdf. Accessed December 16th, 2016.

Walter, C., & Briggs, J. (2012). *What professional development makes the most difference to teachers.* Oxford: University of Oxford Department of Education. Retrieved from http://clie.org.uk/wp-content/uploads/2011/10/Walter_Briggs_2012.pdf. Accessed December 16th, 2016.

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., & Yoon, B. (1995). Teacher reports of opportunity to learn: Analyses of the 1993 California Learning Assessment System. *Educational Evaluation and Policy Analysis, 17*(3), 355-370.

Wood, F. H., & McQuarrie, F. Jr. (1999). On the job learning. New approaches will shape professional learning in the 21st century. *Journal of Staff Development*, *20*, 10-13.

World Bank. (2013). *What matters most for teacher policies: A framework paper* (Systems Approach for Better Education Results Working Paper Series No. 4). Washington, DC: World Bank.

World Bank. (2018). *World Development Report 2018: LEARNING to Realize Education's Promise*. Washington, DC: World Bank.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (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. Retrieved from http://files.eric.ed.gov/fulltext/ED498548.pdf. Accessed December 16th, 2016.

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
Overarching Aspects	64%	78%	27
Content	47%	66%	10
Delivery	42%	69%	14
TOTAL	51%	75%	51
For interviewed programs only:		98%	51

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.

Table 2

Overarching Aspects – Bivariate regressions with robustness checks

Overarching Aspects	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	

$*$ $p < 0.10$, $**$ $p < 0.05$, $***$ $p < 0.01$ correspond to the significance of p-values of robust standard errors. § corresponds to significance at the 10% level or higher for bootstrapped standard errors. † corresponds to significance at the 10% 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	

∗ 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% level or higher for bootstrapped standard errors. †corresponds to significance at the 10% 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).

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	

∗ 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% level or higher for bootstrapped standard errors. † corresponds to significance at the 10% 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 5

Overarching Aspects—Diagnostic table

Overarching Aspects Variables	Top Performers	Obs	At-scale programs	Obs
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

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.

Table 6

Content—Diagnostic table

Content Variables	Top Performers	Obs	At-scale programs	Obs
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

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.

Table 7
Delivery—Diagnostic Table

Delivery Variables	Top Performers	Obs	At-scale programs	Obs
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

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.

Figure 1. Summary of the in-service teacher training survey instrument (ITTSI)

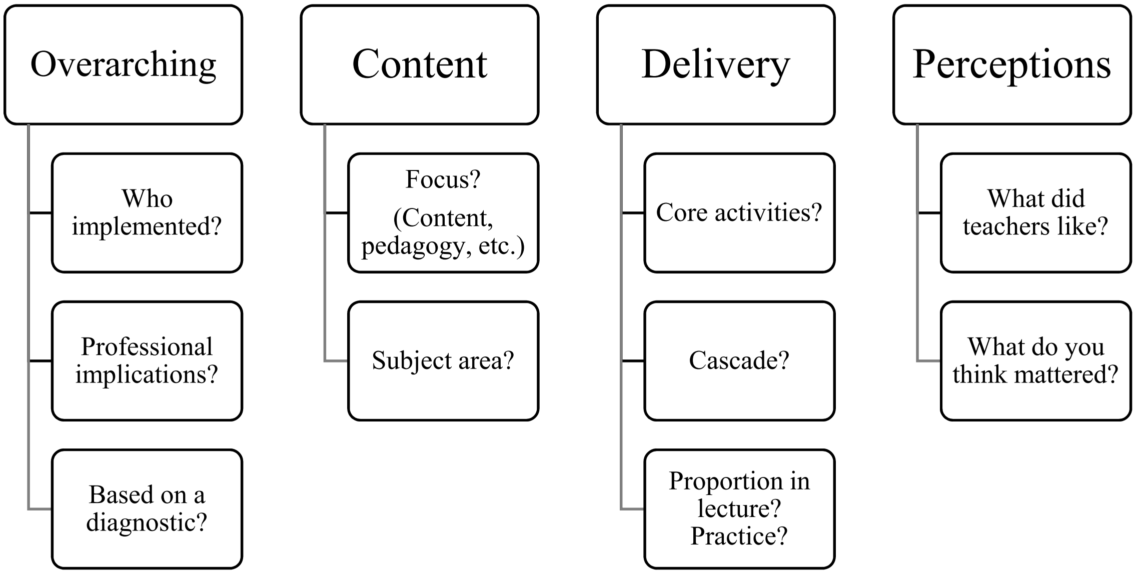


Figure 2. Search process and results for evaluated professional development programs

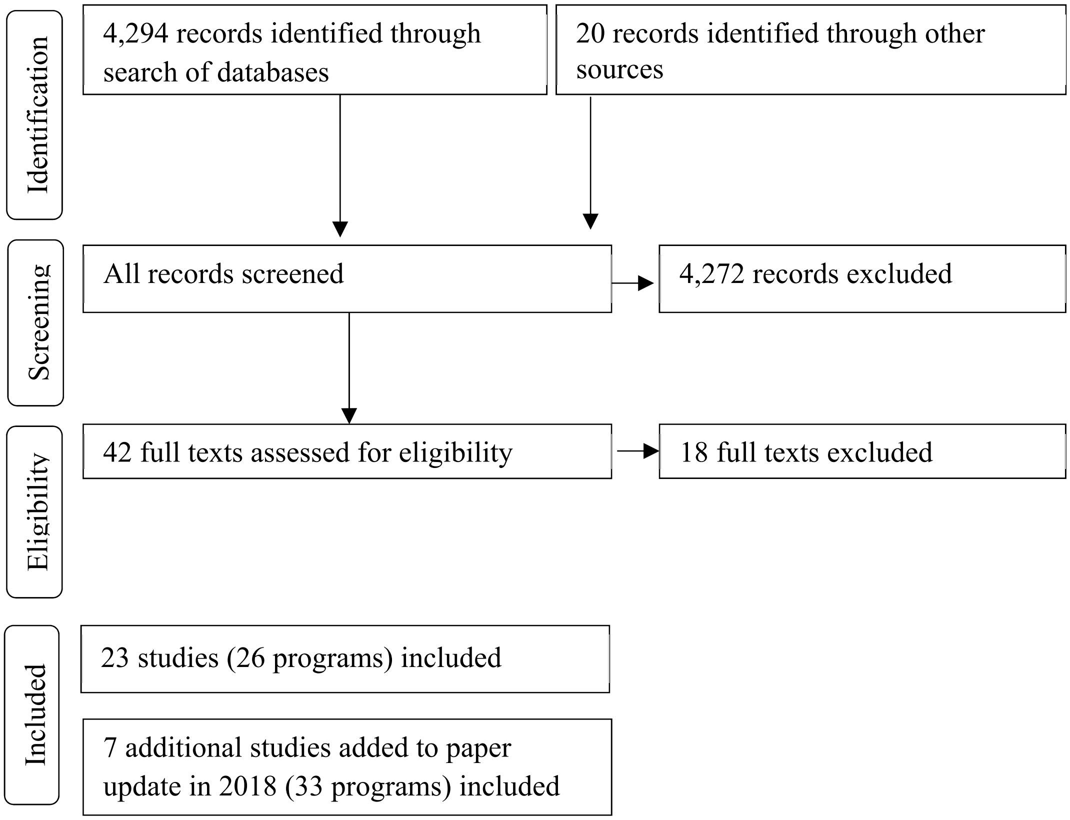
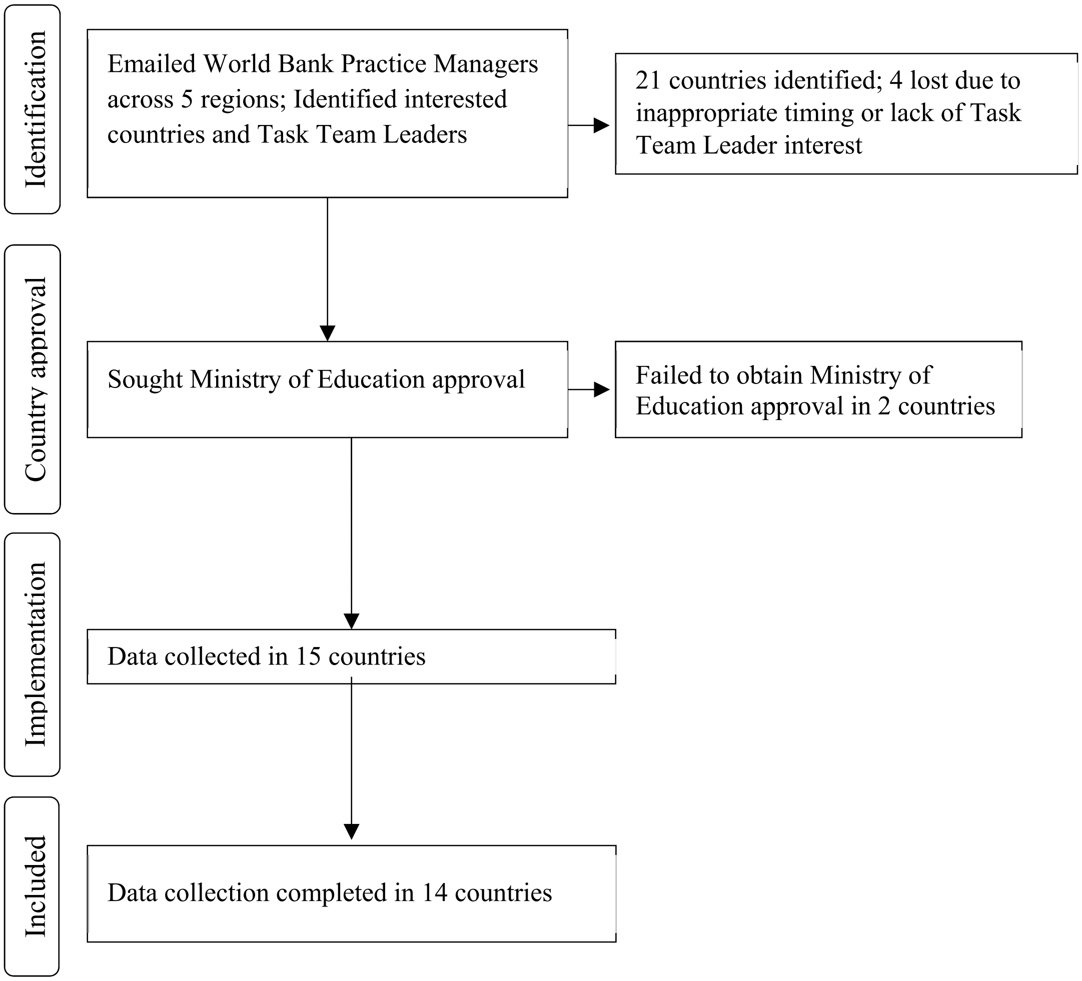

Identification
4,294 records identified through search of databases
20 records identified through other sources
Screening
All records screened
4,272 records excluded
Eligibility
42 full texts assessed for eligibility
18 full texts excluded
Included
23 studies (26 programs) included
7 additional studies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

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

Appendix A: The Survey Instrument

In-Service Teacher Training Survey Instrument (ITTSI)

Introduction

1 What is the name of the in-service teacher training program under discussion?

____

2 What is your full name?

____

3 What is your role in this program?

____

Overarching aspects

4 By the end of this training what is it that you expect teachers to be able to do differently?

____

5 How many years has this program been running?

_____

6 At what scale is this program implemented? (*Please select only one answer)*

1. National □
2. Multiple states or regions □
3. One state or region □
4. Less than one state or region □

7 What kind of organization designed this teacher training program?
(Select all that apply)

1. Government □
2. Non-governmental organization □
3. Private company or social enterprise □
4. Researchers □

8 What kind of organization is implementing this teacher training program?
(Select all that apply)

1. Government □
2. Non-governmental organization □
3. Private company or social enterprise □
4. Researchers □

9	What percentage of the total time teachers spend in this training program detracts from their regular teaching time? ______
10	Is the primary focus of this program teacher training, or is teacher training one part of a broader program? 1. Teacher training is primary focus □ 2. Teacher training is one component □
11	Was the program design based on a diagnostic or evaluation of student learning of some kind? If so, what kind? (*Please select only one answer)* 1. No □ 2. Yes, informal diagnostic □ 3. Yes, formal diagnostic □
12	Was the program design based on a diagnostic or evaluation of teacher skills of some kind? If so, what kind? (*Please select only one answer)* 1. No □ 2. Yes, informal diagnostic □ 3. Yes, formal diagnostic □
13	What teacher skill gaps is this program designed to support? (*Please select only one answer)* 1. Subject content □ 2. Subject-specific pedagogy □ 3. Technology □ 4. Counseling □ 5. Classroom management □ 6. Specific tool □ 7. Assessment □ 8. Curricular update □ 9. General pedagogy □ 10. Theory □
14	Is the program for all teachers or just for certain teachers? 1. All teachers □ 2. Certain teachers □

15	If the program is just for certain teachers, on what characteristics is it targeted? *(Select all that apply)* 1. Geography □ 2. Subject □ 3. Grade □ 4. Teachers' years of experience □ 5. Teachers' skill gaps □ 6. Uncertified Teachers □ 7. Contract teachers □
16	Which grades? *(Select all that apply)* 1. Pre-primary □ 2. Grade 1 □ 3. Grade 2 □ 4. Grade 3 □ 5. Grade 4 □ 6. Grade 5 □ 7. Grade 6 □ 8. Grade 7 □ 9. Grade 8 □ 10. Grade 9 □ 11. Grade 10 □ 12. Grade 11 □ 13. Grade 12 □
17	Are teachers assigned to participate or do they volunteer for the program? 1. Assigned □ 2. Volunteer □ 3. A mix of both □
18	How much do teachers have to pay to register for the program (if anything) per year? _______ A. Amount _______________ B. Noted in what currency?
19	Which of the following other costs do the teachers have to pay to participate in the program? *(Select all that apply)* 1. None □ 2. Transport □ 3. Accommodation □ 4. Materials □ 5. Other □
20	How much do teachers receive as per diem or payment to participate in the program per year? _______ A. Amount _______________ B. Noted in what currency?
21	What is the total cost of the program per year? _______ A. Amount _______________ B. Noted in what currency?

22	Does participation in the training program have any professional implications for teachers? *(Select all that apply)* 1. No □ 2. Status □ 3. Promotion or points towards promotion □ 4. Salary □ 5. Official certification □
23	Are the teachers evaluated at the end of the training? 1. No □ 2. Yes □
24	Is it possible for teachers to fail this exam? 1. No □ 2. Yes □
25	If so, what percentage of teachers fail the exam? ________
26	Is there a positive consequence if teachers are well evaluated? *(Select all that apply)* 1. No □ 2. Status □ 3. Promotion or points towards promotion □ 4. Salary □ 5. Official certification □
27	Is there a negative consequence if teachers are poorly evaluated? *(Select all that apply)* 1. No □ 2. Status □ 3. Promotion or points towards promotion □ 4. Salary □ 5. Official certification □
28	Which of the following are informed about the teachers' performance on the training evaluation? *(Select all that apply)* 1. None □ 2. Teacher □ 3. School where the teacher teaches □ 4. Ministry of Education □

29	What materials, if any, did the program provide alongside the training? *(Select all that apply)* 1. No materials □ 2. Textbooks □ 3. Storybooks or reading pamphlets □ 4. Flashcards □ 5. Word banks □ 6. Computers □ 7. Software □ 8. Teacher manuals □ 9. Lesson plans/videos □ 10. Scripted materials □ 11. Craft materials □
30	How many teachers received training under this program in the last year that the program was implemented? _______
31	In the last year that the program was implemented, what percentage of the teachers who began the training dropped out before the end? _______
32	In how many schools is the program currently being implemented? _______
33	Has this program been evaluated in terms of its impact? 1. No □ 2. Yes □
34	If so, on which of the following was it evaluated in terms of impact? *(Select all that apply)* 1. Teacher knowledge □ 2. Teacher behavior □ 3. Student learning □ 4. Objectives of the program □
35	Over the course of the program, what data are collected centrally? *(Select all that apply)* 1. Frequency of class delivery □ 2. Attendance of participating teachers □ 3. Teachers' assessment of value of training □ 4. Test score of teacher subject knowledge □ 5. Test score of teacher pedagogical knowledge □ 6. Practical test observing teaching □

Content

36 Which of these is the primary focus of the training program? (*Please select only one answer)*

1. Subject content □
2. Subject-specific pedagogy □
3. Technology □
4. Counseling □
5. Classroom management □
6. Specific tool □
7. Assessment □
8. Curricular update □
9. General pedagogy □
10. Theory □

37 Which of these is the secondary focus of the training program? (*Please select only one answer)*

1. No other focus □
2. Subject content □
3. Subject-specific pedagogy □
4. Technology □
5. Counseling □
6. Classroom management □
7. Specific tool □
8. Assessment □
9. Curricular update □
10. General pedagogy □
11. Theory □

38 What is the subject focus of the training program (if any)? *(Select all that apply)*

1. None □
2. Literacy or language □
3. Math □
4. Natural science □
5. Social science □
6. Information technology □
7. Other □

39 Does this program provide training in-person and/or online?

1. In-person □ → ***SKIP TO QUESTION 48***
2. Online □
3. Both □

Online programs

SKIP THIS SECTION FOR PROGRAMS WITH NO ONLINE COMPONENTS

40	In total how many hours of training are provided under this program? ______
41	What proportion of this training do teachers spend practicing with other teachers? ______
42	What proportion of this training do teachers spend practicing with students? ______
43	Over how many weeks is this training spread? ______
44	Do teachers have any contact with a trainer online, as part of the program? 1. No □ 2. Yes □
45	If so, is the contact with trainers individual, in groups, or both? 1. Individual □ 2. Group □ 3. Both □
46	Are the online group sessions compulsory or voluntary? 1. Compulsory □ 2. Voluntary □
47	In total, how many hours of online contact do teachers have with a trainer under the program? _______

Delivery I

48 What are the core activities involved in the training? *(Select all that apply)*

1. Lectures □
2. Discussion □
3. Teaching practice □
4. Discussion of videos □
5. Practice in science labs □
6. Practice with computers □
7. Other practical activities □

49 Which of the following additional activities were included in the training, if any? *(Select all that apply)*

1. None □
2. Development of pedagogical materials □
3. Development of classroom evaluation materials □
4. Training on how to conduct diagnostics □
5. Lesson planning □
6. Using scripted lessons □

50 Does the program use a cascade training model (i.e., program trains trainers who then train teachers)?

1. No □
2. Yes □

51 What is the most common profile of the trainers or facilitators who the teachers have direct contact with? (*Please select only one answer)*

1. Pre-primary, Primary or secondary teacher in the subject of the training □
2. Specially selected expert pre-primary, primary or secondary teacher □
3. Other pre-primary, primary or secondary teacher □
4. University professor or Masters/PhD in education □
5. Researcher □
6. Government official □
7. University student in education □
8. Other □

52	What, if any, training or certification did the trainers or facilitators who the teachers have direct contact with receive? *(Select all that apply)* 1. None □ 2. Designed the program □ 3. Received a specific certification □ 4. Received one week or less of training □ 5. Received more than one week of training □
53	Outside of their normal salary, what kind of engagement mechanisms or incentives are given to trainers? *(Select all that apply)*
	1. None □ 2. Performance related bonus □ 3. Tablet or computer □ 4. Books □ 5. Community recognition □ 6. Other □
54	In total, how many hours of homework are teachers expected to do as part of the training, per year? ______
55	Over how many weeks is this homework spread? ______
56	Which of these types of follow-up support do teachers receive? *(Select all that apply)* 1. Text messages □ 2. Phone calls □ 3. Emails □ 4. In-school support from principals □ 5. In-school support from other school staff □
57	Over how many weeks is this follow-up support spread? ______
58	Does the program provide any face-to-face training? 1. No □ → ***SKIP TO QUESTION 71*** 2. Yes □

Delivery II

SKIP THIS SECTION FOR PROGRAMS WITH NO FACE-TO-FACE COMPONENTS (I.E. ONLINE ONLY)

59	How many days do teachers work face-to-face with trainers or facilitators in this program? ______
60	Over how many weeks is this face-to-face training spread? ______
61	Approximately what proportion of this time is spent in lectures and discussion? ______
62	Approximately what proportion of this time is spent “practicing teaching” with *students*? ______
63	Approximately what proportion of this time is spent “practicing teaching” with *other teachers*? ______
64	Approximately what proportion of this time is spent in other practical activities with other teachers? ______
65	Where does the majority of the face-to-face training take place? (*Please select only one answer)* 1. School of teacher being trained □ 2. Central location (other school, hotel, government building etc.) □ 3. University or training center □
66	On average, about how many teachers are there per trainer or facilitator in each training session? ______
67	How many in-school follow-up support visits do teachers receive after the initial training (if any)? ______
68	What is the nature of these follow-up visits? *(Select all that apply)* 1. In-class pedagogical support □ 2. Monitoring □ 3. Review material □

69	Over how many weeks are the follow-up visits spread? _______
70	How many times do teachers receive any of the above types of support? *(Count each text message/phone call/conversation as one time.)*
71	What is the total duration of this program in days? _______

Delivery III

72	Were there any elements of the program that the teachers particularly *liked*? Element 1 ____________________ Element 2 ____________________ Element 3 ____________________
73	Were there any elements of the program that the teachers particularly *disliked*? Element 1 ____________________ Element 2 ____________________ Element 3 ____________________
74	What were the key elements you think made the program work? Element 1 ____________________ Element 2 ____________________ Element 3 ____________________

Appendix B: The List of Included Papers

Abeberese, A. B., Kumler, T. J., & Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. *Journal of Human Resources*, *49*(3), 611-633. doi:10.1353/jhr.2014.0020

Angrist, J., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics, 19*, 343-69. doi:10.1086/319564

Bando, R. & Li, X. (2014). The effect of in-service teacher training on student learning of English as a second language. Washington, DC: Inter-American Development Bank.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy*, *2*(1), 1-30. doi:10.1257/pol.2.1.1

Banerjee, A., Cole, S., Duflo, E., & Linden, L. L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics, 122*, 1235-1264. doi:10.1162/qjec.122.3.1235

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia* (World Bank Policy Research Working Paper No. 4836). Washington, DC: World Bank. doi:10.1596/1813-9450-4836

Berlinski, S. & Busso, M. (2015). Challenges in Educational Reform: An Experiment on Active Learning in Mathematics (IDB Working Paper Series No. IDB-WP-561). Washington, DC: Intera-American Development Bank.

Beuermann, D. W., Naslund-Hadley, E., Ruprah, I. J., & Thompson, J. (2013). The pedagogy of science and environment: Experimental evidence from Peru. *The Journal of Development Studies*, *49*(5), 719-736. doi:10.1080/00220388.2012.754432

Brooker, S., Inyega, H., Estambale, B., Njagi, K., Juma, E., Jones, C., et al. (2013). *Impact of malaria control and enhanced literacy instruction on educational outcomes among Kenyan school children: A multi-sectoral, prospective, randomized evaluation*. 3ie Draft Grantee Final Report. Washington, DC: International Initiative for Impact Evaluation. Retrieved from http://erepository.uonbi.ac.ke/bitstream/handle/11295/81003/Brooker_Impact of malaria control and enhanced literacy.pdf?sequence=1&isAllowed=y. Accessed December 16th, 2016.

Carillo, P., Onofa, M., & Ponce, J. (2010). *Information technology and student achievement: Evidence from a randomized experiment in Ecuador* (IDB Working Paper No. 223). Washington, DC: Inter-American Development Bank.

Cilliers, J., Fleish B., Prinsloo C., & Taylor, S. (2018). How to improve teaching practice? Experimental comparison of centralized training and in-classroom coaching. Unpublished manuscript.

Facundo Albornoz, Anauati, M., Furman, M., Luzuriaga, M., Podestá, M., & Taylor, I. (2017).Training to teach science: experimental evidence from Argentina. Discussion Papers 2017-08, University of Nottingham, CREDIT. Retrieved from https://ideas.repec.org/p/not/notcre/17-08.html. Accssed August 1st, 2018.

He, F., Linden, L. L., & MacLeod, M. (2008). *How to teach English in India: Testing the relative productivity of instruction methods with Pratham English Language Education Program.* Unpublished manuscript, Columbia University, New York, NY.

Kerwin, J. T., & Thornton, R. (2015). *Making the grade: Understanding what works for teaching literacy in rural Uganda*. Unpublished manuscript, University of Michigan, Ann Arbor, MI.

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2012). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing* (REAP Working Paper No. 228). Stanford, CA: Stanford University, Rural Education Action Project.

Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., et al. (2012). *Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in public schools in rural minority areas in Qinghai* (REAP Working Paper No. 237). Stanford, CA: Stanford University, Rural Education Action Project.

Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., et al. (2013). Can information and counseling help students from poor rural areas go to high school? Evidence from China. *Journal of Comparative Economics*, *41*, 1012-1025. doi:10.1016/j.jce.2013.06.004

Lucas, A. M., & Mbiti, I. M. (2012). Access, sorting, and achievement: The short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, *4*(4), 226–225. doi:10.1257/app.4.4.226

Mo, D., Zhang, L., Lui, R., Qu, Q., Huang, W., Wang, J., et al. (2013). *Integrating computer-assisted learning into a regular curriculum: Evidence from a randomized experiment in rural schools in Shaanxi* (REAP Working Paper No. 248). Stanford, CA: Stanford University, Rural Education Action Project.

Nitsaisook, M., & Anderson, L.W. (1989). An experimental investigation of the effectiveness of inservice teacher education in Thailand. *Teaching & Teacher Education*, *5*(4), 287-302. doi:10.1016/0742-051X(89)90027-9

Piper, B., & Korda, M. (2011). *EGRA Plus: Liberia* (Program evaluation report). Durham, NC: RTI International. Retrieved from http://files.eric.ed.gov/fulltext/ED516080.pdf. Accessed December 16th, 2016.

Piper, B. & Zuilkowski, S. (2017). Teacher coaching in Kenya: Examining instructional support in public and nonformal schools. *Teaching & Teacher Education*, 47, 173-183.

Piper, B., Zuilkowski, S., & Ong'ele, S. (2016). Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya. *Comparative Education Review, 60(4), 776-807.*

Pournara, C., Hodgen, J., Adler, J., & Pillay, V. (2015). Can improving teachers' knowledge of mathematics lead to gains in learners' attainment in Mathematics?. *South African Journal of Education*, *35*(3), 1-10. doi:10.15700/saje.v35n3a1083

Spratt, J., S. King, & Bulat, J. (2013). Independent Evaluation of the Effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) Program in Mali. Menlo Park, CA: The William and Flora Hewlett Foundation. Retrieved from http://www.educationinnovations.org/sites/default/files/Mali RLL evaluation Endline Report - REVISED 30nov2013 (6).pdf. Accessed March 10, 2017.

Tan, J., Lane, J., & Lassibille, G. (1999). Student outcomes in Philippine elementary schools: An evaluation of four experiments. *World Bank Economic Review*, *13*, 493-508. doi:10.1093/wber/13.3.493

Weir, C. J. , & Roberts, J. (1991). Evaluating a teacher Training project in difficult circumstances. In S. Anivan (Ed.), *Issues in Language Programme Evaluation in the*

1990's. Anthology serires 27, 91-109. Singapore: Southeast Asian Ministers of Education Organization.

Yue, A., Shi, Y., Chang,F., Yang, C., Wang, H., Yi,H., Luo, R., Liu,C., Zhang, L., Chu, J.,&Rozelle, S. (2013). Dormitory management and boarding students in China's rural elementary schools. *China Agricultural Economic Review*, 6(3), 523-550.

Zhang, D., & Campbell, T. (2012). An exploration of the potential impact of the integrated experiential learning curriculum in Beijing, China. *International Journal of Science Education*, *34*(7), 1093-1123. doi:10.1080/09500693.2011.625057

Zhang, L., Lai, F., Pang, X., Yi, H., & Rozelle, S. (2013). The impact of teacher training on teacher and student outcomes: evidence from a randomised experiment in Beijing migrant schools. *Journal of development effectiveness*, *5*(3), 339-358. doi:10.1080/19439342.2013.807862

Appendix C: Mathematical Appendix

For all estimates included in the meta-analysis, the goal is to estimate the standardized effect size, again drawing on Borenstein et al. (2009):

$$d = \frac{D}{S_{pooled}} \qquad \text{(Equation 1)}$$

using some estimate of the raw mean difference between treatment and control groups, *D*, as well as its combined standard deviation for treatment and control groups, *S_{pooled}*. All studies report *D* directly, however, *S_{pooled}* is commonly not reported. Almost all studies we review instead report the standard error of *D, SE_D.* Where this is the case, if we assume that the standard deviations of the two groups are the same, then the variance of D is:

$$V_D = \frac{n_1 + n_2}{n_1\, n_2}\, S_{pooled}^2 \qquad \text{(Equation A1)}$$

where *n_1* and *n_2* are the sample sizes in the two groups. The standard error of *D* is then the square root of *V*.

$$SE_D = \sqrt{V_D} \qquad \text{(Equation A2)}$$

Combining Equation A1 and Equation A2, we derive our equation for *S_{pooled}*, the within-groups standard deviation, pooled across treatment and control groups:

$$SE_D = \sqrt{\frac{n_1 + n_2}{n_1\, n_2}\, S_{pooled}^2}$$

$$SE_D = \sqrt{\frac{n_1 + n_2}{n_1\, n_2}}\sqrt{S_{pooled}^2}$$

$$SE_D = \sqrt{\frac{n_1 + n_2}{n_1\, n_2}} S_{pooled}$$

$$S_{pooled} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, SE_D \qquad \text{(Equation 2)}$$

We can then divide *D* by *S_{pooled}* to calculate standardized effect sizes, *d*, for all estimates.

Appendix D: Full List of Characteristics of At-Scale Programs and Evaluated Programs

Table D1

Overarching Aspects – Descriptive Statistics

	All Evaluated Programs			At-scale Programs		
Overarching Aspects	Mean	Standard Deviation	Obs	Mean	Standard Deviation	Obs
Designed by Government	0.152	0.364	33	0.795	0.405	132
Designed by NGO or social enterprise	0.394	0.496	33	0.068	0.253	132
Designed by researchers	0.424	0.502	33	0.136	0.344	132
Implemented by Government	0.273	0.452	33	0.896	0.309	48
Implemented by NGO or social enterprise	0.394	0.496	33	0.250	0.438	48
Implemented by researchers	0.333	0.479	33	0.125	0.334	48
Design not based on diagnostic	0.121	0.331	33	0.479	0.505	48
Design based on informal diagnostic	0.242	0.435	33	0.383	0.491	47
Design based on formal diagnostic	0.333	0.479	33	0.563	0.501	48
Targeting by geography	0.533	0.507	30	0.125	0.334	48
Targeting by subject	0.300	0.466	30	0.188	0.394	48
Targeting by grade	0.806	0.402	31	0.313	0.468	48
Targeting by years of experience	0.067	0.254	30	0.125	0.334	48
Targeting by skill gaps	0.033	0.183	30	0.208	0.410	48
Targeting by contract teachers	0.100	0.305	30	0.063	0.245	48
Participation has no implications for teacher status, salary, or promotion	0.364	0.489	33	0.417	0.498	48
Participation has status implications only	0.061	0.242	33	0.250	0.438	48
Participation has implications for salary or promotion	0.303	0.467	33	0.250	0.438	48
Teachers are not evaluated	0.212	0.415	33	0.563	0.501	48
Positive consequence if teachers are well evaluated	0.121	0.331	33	0.375	0.489	48
Negative consequence if teachers are poorly evaluated	0.061	0.242	33	0.167	0.377	48
Program provides materials	0.867	0.346	30	0.958	0.202	48
Program provides textbooks	0.214	0.418	28	0.292	0.459	48
Program provides storybooks	0.321	0.476	28	0.125	0.334	48
Program provides computers	0.143	0.356	28	0.125	0.334	48
Program provides teacher manuals	0.552	0.506	29	0.625	0.489	48
Program provides lesson plans/videos	0.321	0.476	28	0.542	0.504	48
Program provides scripted lessons	0.241	0.435	29	0.333	0.476	48
Program provides craft materials	0.107	0.315	28	0.333	0.476	48
Program provides other reading materials (flashcards, word banks, reading pamphlets)	0.357	0.488	28	0.208	0.410	48
Program provides software	0.276	0.455	29	0.188	0.394	48
Number of teachers trained	655.7	1,514.9	19	8,514.5	37,582.2	139
Number of schools in program	95.6	149.5	28	6,367.3	18,281.7	29
Program age (years)	2.6	2.8	25	3.8	4.5	138
Dropouts in last year	0.5	0.5	15	58.9	346.9	43

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.

Table D2

Content – Descriptive Statistics

	All Evaluated Programs			At-scale Programs		
Content	Mean	Standard Deviation	Obs	Mean	Standard Deviation	Obs
Focus is subject content	0.636	0.489	33	0.273	0.447	139
Focus is pedagogy	0.576	0.502	33	0.374	0.486	139
Focus is technology	0.212	0.415	33	0.137	0.345	139
Focus is counseling	0.091	0.292	33	0.036	0.187	139
Focus is classroom management	0.121	0.331	33	0.079	0.271	139
Focus is a specific tool	0.091	0.292	33	0.065	0.247	139
No subject focus	0.061	0.242	33	0.083	0.279	48
Subject focus is literacy/language	0.515	0.508	33	0.521	0.505	48
Subject focus is math	0.152	0.364	33	0.542	0.504	48
Subject focus is science	0.091	0.292	33	0.292	0.459	48
Subject focus is information technology	0.030	0.174	33	0.229	0.425	48
Subject focus is language & math	0.061	0.242	33	0.000	0.000	48
Subject focus is other	0.030	0.174	33	0.229	0.425	48
Training involves lectures	0.950	0.224	20	0.604	0.491	139
Training involves discussion	0.750	0.444	20	0.842	0.366	139
Training involves lesson enactment	0.600	0.503	20	0.727	0.447	139
Training involves materials development	0.200	0.410	20	0.729	0.449	48
Training involves how to conduct diagnostics	0.238	0.436	21	0.354	0.483	48
Training involves lesson planning	0.480	0.510	25	0.625	0.489	48
Training involves use of scripted lessons	0.333	0.482	24	0.438	0.501	48

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.

Table D3

Delivery – Descriptive Statistics

	All Evaluated Programs			At-scale Programs		
Delivery	Mean	Standard Deviation	Obs	Mean	Standard Deviation	Obs
Cascade training model	0.519	0.509	27	0.587	0.494	138
Trainers are primary or secondary teachers	0.152	0.364	33	0.416	0.495	137
Trainers are experts - university professors / graduate degrees in education	0.212	0.415	33	0.560	0.499	84
Trainers are researchers	0.091	0.292	33	0.051	0.221	137
Trainers are local government officials	0.242	0.435	33	0.022	0.147	137
Trainers are education university students	0.030	0.174	33	0.000	0.000	139
Initial period of face-to-face training for several days in a row	0.938	0.246	32	0.854	0.357	48
Total hours of face-to-face training	59.742	39.667	31	13.265	14.864	34
Proportion of face-to-face training spent in lectures	0.534	0.290	17	0.481	0.296	35
Proportion of face-to-face training spent practicing with students	0.071	0.107	19	0.082	0.154	36
Proportion of face-to-face training spent practicing with teachers	0.376	0.341	19	0.156	0.183	34
Duration of program (weeks)	9.800	13.566	30	7.216	12.559	37
Training held at schools	0.030	0.174	33	0.108	0.315	37
Training held at central location including hotel conference room etc.	0.576	0.502	33	0.730	0.450	37
Training held at university or training center	0.091	0.292	33	0.162	0.374	37
Number of teachers per training session	30.794	10.227	17	30.972	15.157	36
Includes follow-up visits	0.760	0.436	25	0.496	0.502	139
Follow-up visits for in-class pedagogical support	0.333	0.479	33	0.375	0.489	48
Follow-up visits for monitoring	0.242	0.435	33	0.333	0.476	48
Follow-up visits to review material	0.091	0.292	33	0.104	0.309	48
Includes distance learning	0.167	0.381	24	NA		
Duration of distance learning (months)	8.556	8.763	27	NA		

Obs refers to the number of PD programs in each sample (top performing, evaluated programs, and at-scale programs) that report whether or not they have a given characteristic.