WPS7834

Policy Research Working Paper 7834

Background Paper to the 2018 World Development Report

# Training Teachers on the Job

## What Works and How to Measure It

*Anna Popova*
*David K. Evans*
*Violeta Arancibia*

WORLD BANK GROUP
Africa Region
Office of the Chief Economist
September 2016

Policy Research Working Paper 7834


## Abstract

A significant body of research demonstrates that teachers and the quality of their teaching are crucial components of student learning. Many teachers in resource-poor environments have limited knowledge, skills, or motivation. Some impact evaluations have shown promising results from interventions to improve the quality of teaching. This paper reviews the existing body of evidence on what kinds of in-service teacher training interventions are most effective, and highlights the knowledge gaps. It reveals the dearth of detail on the nature of teacher training interventions and proposes a standard set of indicators—the In-Service Teacher Training Survey Instrument—for reporting on such programs as a prerequisite for understanding which interventions lead to improved student learning. Across a set of 26 programs with impact evaluations and student learning results, programs that provide complementary materials, focus on a specific subject, and include follow-up visits tend to show higher gains. Programs that use non-education professionals as trainers tend to have worse outcomes. Statistical power to identify these effects is limited, and use of these standard indicators in future impact evaluations will facilitate more precise inference.


---

This paper—prepared as a background paper to the World Bank's *World Development Report 2018: Realizing the Promise of Education for Development*—is a product of Office of the Chief Economist, Africa Region. 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://econ.worldbank.org. The authors may be contacted at apopova@stanford.edu and devans2@worldbank.org.

# Training Teachers on the Job:
# What Works and How to Measure It

Anna Popova David K. Evans 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, Barbara Bruns, Joost de Laat, Margaret Dubeck, Deon Filmer, Andrew Ragatz, and Halsey Rogers, and for financial support from the Strategic Impact Evaluation Fund at the World Bank. They are also grateful for excellent research assistance by Fei Yuan, for the generous time provided by numerous teacher training implementers to fill in knowledge gaps, and for data on World Bank projects from Mary Breeding. Most work for this paper was completed when Popova and Evans worked in the Office of the Chief Economist for the Africa Region at the World Bank, and Arancibia is based at the Catholic University of Chile. The corresponding authors can be contacted at Popova (apopova@stanford.edu) and Evans (devans2@worldbank.org).

## 1. Introduction

Increases in access to education across the developing world in recent years have stimulated a shift in policy focus towards improving education quality, often measured by student test score gains. Controlling for socioeconomic factors, teachers have been argued to be the most important determinant of student learning. The difference between a weak teacher and a great teacher has been measured – in the United States – at up to a full year of student learning (Hanushek and Rivkin 2010). Similarly, in Chile, the impact of good teachers over multiple years accumulates to the equivalent of 0.3 standard deviation (SD) in secondary school (MINEDUC 2009). Beyond these immediate improvements in student learning, teachers who raise student test scores also significantly improve students' long-term outcomes, such as their probability of graduating college and adult salaries, while decreasing the likelihood of teenage pregnancy (Chetty et al. 2014).

Furthermore, in a recent review 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 – was among the most recommended interventions for improving student learning (Evans and Popova, forthcoming; Kremer, Brannen, and 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 by 0.79 SD (Piper & Korda 2011). Similarly, a program in Kenya that streamed students into classes based on ability so that teachers could tailor teaching to the appropriate level increased test scores by 0.17 SD in language and 0.16 SD in math (Kremer, Duflo, & Dupas 2011).

In-service teacher training – or 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. Massive amounts of government 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 training 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.

While mechanisms that improve teacher practice at pre-service are important, this paper focuses on in-service training, motivated partly by the fact that it faces fewer institutional constraints to change in many countries. In other words, it is often easier for a government to implement an innovative in-service teacher training program than to reform pre-service education, since the latter relies on more institutions outside the Ministry of Education. Furthermore, in-service teacher training policies tend to receive support from teachers' unions, and provide an opportunity to improve the quality of both existing teachers (the stock) as well as future teachers (the flow).

Nonetheless, there is a sizeable evidence gap when it comes to whether resources spent on in-service teacher training are improving learning, partly owing to a lack of instruments designed to measure teacher professional development. Instruments exist to capture the design of teacher policy – SABER Teachers (World Bank 2013) – on one end of the policy-practice spectrum, and how teachers behave in the classroom – Stallings (Stallings 1977), the Classroom Assessment Scoring System or CLASS (La Paro & Pianta 2003), and others – at the other end of the spectrum. 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, forthcoming; McEwan 2014); in other words, teacher training programs vary enormously, both in their form and in their effectiveness.

This paper has two objectives. The first is to fill this gap in information on the essential characteristics of training 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. The instrument was piloted on a sample of programs from low- and middle-income countries whose impact has already been evaluated, and the resulting data analyzed 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, seeking to provide a richer analysis than the more general analysis – on education interventions broadly – in Evans and Popova (forthcoming).

Findings suggest that characteristics positively associated with program impact on student learning include the provision of textbooks and other reading materials alongside the training, as well as linking participation to incentives such as promotion or salary implications, and the training having a specific subject focus, 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 responsible for positive impacts on student learning.

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 search strategy, the instrument design, the data collection, and the analytical strategy. Section 4 presents the results of our quantitative and qualitative analyses, and Section 5 concludes.

## 2. Background

### 2.1. Theory

The education literature suggests a range of factors to consider in the design of teacher training programs, sometimes illustrated by empirical work. A brief discussion of this literature follows, organized around six questions.

*Who is learning?* Because teachers are the students in in-service teacher training, principles of adult education are relevant. Adult education tends to work best with clear applications rather than a theoretical focus (Cardemil 2001; Knowles et al. 2005). Teacher training will work best if it adjusts to different points in the teachers' careers, i.e., 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).

*Who is teaching?* Unsurprisingly, the quality of trainers is crucial to teacher learning, just as the quality of teachers is crucial to student learning (Knowles et al. 2005). This calls into question the standard cascade model of training in low-income environments, in which both information and pedagogical ability may be diluted as a master trainer trains a trainer, and so forth.

*How to train?* At least two key characteristics emerge from the literature on how to train. First, teachers should learn how to carry out formative evaluation so that they can effectively evaluate their own progress towards their teaching goals (Bourgeois and Nizet 1997); second, teacher training will work best with concrete, realistic goals (Baker and Smith 1999).

*How long to train?* There is no theoretical consensus on exactly how long training should last, although there is suggestive evidence in the literature, to be discussed in the next section. In short, there is no reason to believe that one-off workshops will be effective, despite their being common in school systems.

*What to teach?* 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, and teachers need training in different pedagogies for these methods.

*Where to teach?* In general, teacher training 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 (Woods and McQuarrie 1999). However, this will depend on the environment. In very difficult teaching environments, some degree of training outside the school may facilitate focus on the part of the trainees (Kraft and Papay 2014).

### 2.2. What works in high-income countries?

While a full review of the literature in high-income countries is beyond the scope of this study, it may be useful to highlight recent work on in-service teacher training 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 from the U.S. that measure student test scores as an outcome examined the impact of both "general" and "managed" professional development, relative to other interventions (Fryer 2016). General professional development (PD) – as the name suggests – leaves a fair amount of flexibility, even while 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 was found to increase 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 7 general PD studies and 2 managed PD studies.

Another U.S.-focused review found that professional development 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 9 were identified as having 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 5 of those were high-quality randomized trials. As the authors conclude, "The limited research on effectiveness means that schools and districts cannot use evidence of effectiveness alone to narrow their choice" (Gersen 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. Characteristics of effective programs included that (1) they were not a "one-shot workshop" but

rather embedded in the curriculum, and (2) they were also “sustained” and “intense” (Wei et al. 2009).

Much evidence from other high-income countries is also more qualitative. Narrative empirical analysis by Darling-Hammond et al. (2010) highlights that in high-achieving countries, in-service support to teachers includes (1) mentoring for all beginners, coupled with a reduced teaching load and shared planning time for new and mentor teachers, (2) extensive opportunities for ongoing professional learning, embedded in substantial planning and collaboration time at school, and (3) 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 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 experimental or high-quality quasi-experimental evidence makes it difficult to draw detailed conclusions about what works within teacher training even in the rich country literature. However, from a combination of the more rigorous quantitative and qualitative studies above, there is suggestive evidence that in-service teacher training 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. Methods

### 3.1. Search Strategy

We searched the existing literature on in-service teacher training in developing countries to identify a sample of training programs which 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 training in developing countries. Secondly, we would use 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 are impact evaluations of primary education interventions in low- and middle-income countries that either (1) focused on in-service teacher training, or included this as one component of a broader program, and (2) which reported impacts of the program on student test scores in math, language, or science. We include 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 training from the sample of a review of education reviews that two of the authors conducted last year (Evans and Popova, forthcoming). We examined the 4,314 results from both sources to exclude articles that from their title and blurb, 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 the final 23 papers, evaluating 26 programs between them. The search process is detailed in Figure 1. The 23 papers are listed in Appendix A.

### 3.2. The In-Service Teacher Training Survey Instrument (ITTSI)

The ITTSI was designed based on (1) the descriptive, impact evaluation, and theoretical literatures characterized above, and (2) the authors' previous 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. The ITTSI is summarized in Figure 2.

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 it's 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), 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), 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 program effectiveness, such as 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 classify the full range of programs and the dimensions along which they differ. This resulted in a draft instrument consisting of a total of 51 indicators, for which we collected and analyzed data, the results of which are reported in this paper. Subsequent to this data collection and analysis, we shared our results with a series of experts 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 3 pieces of meta-data, is presented in Appendix B.

### 3.3. Data Collection

Data collection and coding for our sample of 26 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 (50%) 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). While 60% of details concerning overarching aspects of teacher training programs – such as whether the program was designed by a government or non-governmental organization (NGO) – can be extracted from the evaluations, on average, only 43% and 36% 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 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 17 of the 26 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 12 of the 26 programs. In four 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. The interviews loosely followed the survey

instrument, while including open-ended questions and space for program implementers to provide any additional program information which they perceived as important.

For the 12 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 80% for Overarching Aspects indicators, 67% of Content indicators, and 68% of Delivery indicators (Table 1).

### 3.4. Analytical Strategy

For our sample of in-service teacher training 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 consists of estimating a bivariate linear regression of the form:

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

where $Y$ is the standardized impact on student test scores, $X$ an indicator of a given program characteristic, $\alpha$ is the constant, $\beta$ is the coefficient on $X$, $\varepsilon$ is the estimation error, and $i$ represents each in-service teacher training program in the sample.

In preparation for this analysis, we standardize the impact estimates, $Y$, for each of the programs. We convert the independent program variables, $X$, to dummy variables wherever possible to facilitate comparability of coefficients.

First, while 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. 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 training program is compared to a control group taught by teachers who did not receive said training. Almost all the studies in our sample use difference-in-differences methods to estimate the effect of the teacher training programs

– or the larger programs of which training is a sub-component – on student learning and report 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.

Turning to the independent variables, as originally coded, the 51 indicators for which we collected information capturing various design and implementation characteristics of the training programs took a number of forms. These consisted of dummy 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 dummy variables. For categorical variables, this is straightforward. For example, we convert the original categorical variable for the location of the initial teacher training – 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.

We then conduct bivariate regressions on this set of complete dummy variables with continuous impact estimates on test scores as our dependent variable for each regression. We supplement this regression analysis with a qualitative analysis of what works, relying on the self-reported perceptions of program implementers along three dimensions: (1) which program elements they identified as most responsible for any positive impacts on student learning, (2) which elements, if any, teachers particularly liked, and (3) which elements, if any, teachers particularly disliked.

## 4. Results: What do teacher training programs look like, and what training program characteristics are associated with better test score gains and higher satisfaction among teachers?

This section contributes to the thus far limited knowledge base on the specific elements of teacher training programs that successfully improve student learning in low- and middle-income countries by presenting the results of our quantitative and qualitative analyses, in turn, for our sample of all programs whose impact has been evaluated. We first present descriptive statistics for all program characteristics for our sample of 26 programs, before proceeding to the results of our quantitative and qualitative analyses. It is worth remembering that these are characteristics of programs that have been rigorously evaluated, which are likely very different from the universe of teacher training programs. For example, one might imagine that the programs that are rigorously evaluated are those that researchers expected to have the highest probability of being effective, or those that are small-enough scale to permit a comparison group.

We then present associations between gains in student learning and characteristics of teacher training programs. With 26 observations, the statistical power to detect any significant impacts is extremely low, so this exercise is principally useful for observing a pattern of suggestive results rather than drawing any definitive conclusions. If future impact evaluations gather a consistent set of indicators – using, for example, the ITTSI, then more conclusive analysis will be possible at a later date.

### 4.1. What do evaluated teacher training programs look like?

For each of our quantifiable instrument categories - Overarching Aspects, Content, and Delivery – we calculate means and standard deviations for all indicators so as to describe, in detail, what programs in our sample look like, before analyzing which of these program characteristics are conducive to improving student learning. In what follows we summarize some of the more common program traits across the sample; full descriptive statistics can be found in Tables 2-4. Taking first the Overarching Aspects of the training programs in our sample (Table 2), these programs are most commonly designed by researchers (46%) and based on some kind of formal diagnostic or evaluation (41%). Consistent with the fact that most evaluated programs are pilot programs run by researchers or non-government organizations (73% combined), they tend to be small in scale with an average of 609 teachers receiving training across 57 schools, per year. The majority of programs target beneficiary teachers by the grade they teach (79%), and some target teachers by subject (25%), but none targets teachers based on their years of experience or specific skill gaps. Interestingly, for a large proportion of programs (41%), participation does have implications for salary or promotion.

Turning next to Content-related program characteristics (Table 3), the most common primary focus of the programs in our sample is on pedagogy (46%) - as opposed to technology or classroom management, etc. – and the most common secondary focus is on subject content (68%). Within subjects, the majority of programs focus on either literacy/language, or math, or both (90%). Most training programs are also linked with some sort of materials provision – e.g., textbooks, storybooks, teacher manuals, lesson plans – (82%), and involve lectures (93%), discussion (64%), and lesson enactment (57%), with fewer including the use of scripted lessons (33%) or training on how to conduct diagnostics of student performance (33%).

Finally, in terms of Delivery characteristics (Table 4), almost all of the programs in our sample have teachers meet face-to-face for some number of consecutive days of initial training (92%), and in most cases this training takes place at a central location such as a local government building or hotel conference room (82%), as opposed to in-school (6%). For half the programs this training is delivered using a cascade model. The majority of programs also include some kind of follow-up support visits (78%), with an average of 6 visits received per teacher. These visits are mixed between monitoring (47%) and in-class pedagogical support (41%), and are less often used for reviewing

material (12%). Overall, programs provide on average 64 hours of training in total, with 48% of this time dedicated to lectures, 52% involving practice with other teachers, and only 6% geared towards practicing with students.

### 4.2.Quantitative Analysis

We discuss, for each of our quantifiable instrument categories - Overarching Aspects, Content, and Delivery – those variables we observe to be most associated with student learning gains. Tables 5, 6, and 7 present the results of our bivariate regressions for each of these categories in turn, in each case reporting the results for the five variables displaying the largest association with program impact in absolute value, in descending order. The complete set of regression results for all variables included in our instrument, similarly presented by category, are included in Appendix D.

Among Overarching Aspects (Table 5), the largest impacts on student learning come from the provision of certain materials alongside the teacher training. In this sample, providing textbooks alongside training is associated with a test score gain of 0.36 standard deviation (significant with 95% confidence), the provision of storybooks is associated with a gain of 0.13 standard deviation (not significant), and providing other reading materials increases the likelihood of improving student learning by 0.16 standard deviation (significant at 90%). The only other characteristic – given our small sample – with a statistically significant association relates to incentives. When participation in a teacher training programs has explicit implications for promotion, salary increases, or allows teachers to earn points that contribute to either of these, test scores are 0.14 standard deviation higher (significant at 90%). Targeting participant teachers by their years of experience, while insignificant, has the next largest association with student learning, also at 0.14 standard deviation. However, this is entirely driven by a single program, the Balsakhi program in rural India which trains women from the local community who have completed secondary school to provide remedial education to students falling behind (Banerjee et al. 2007). Indeed, that is the only program out of the 26 that explicitly targeted teachers based on their experience.

Among the Content variables (Table 6), a primary focus on classroom management is associated with the largest gain in test scores, with those programs showing 0.47 standard deviations greater impact on student learning. But this again is driven by a single program, the In-Service Teacher

Education program in Thailand (Nitsaisook & Anderson 1989), which is a cluster randomized-control trial but with few clusters. In contrast, in general, training programs that are not focused on a given subject – i.e., those that are focused on pedagogy, new technology, classroom management, assessment, or counseling, without direct application to math, language or another subject – are associated with 0.24 lower standard deviation in student learning. Meanwhile, programs which have a primary focus on pedagogy or training, and those that have a secondary focus on subject content are both associated with 0.18 standard deviation larger impacts on student learning. Likely due to limited statistical power, none of these associations are statistically significant.

Turning to Delivery characteristics (Table 7), the highest association with student learning gains is holding face-to-face training in a university or training center (as opposed to a central location such as a hotel or government administrative building, which was the omitted category), which is associated with 0.39 higher standard deviation in student learning (significant at 90%). 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 – improves program impact on student learning by 0.26 standard deviation. The profile of teacher trainers seems important, with using researchers or local government officials – as opposed to education practitioners of some sort – as the trainers in direct contact with teachers being associated with 0.20 and 0.17 standard deviation lower program impacts on student test scores, respectively. Finally – and in alignment with recent literature highlighting the overly theoretical nature of many training programs as an explanation for their limited effects on student learning – the proportion of training time spent practicing with other teachers is highly correlated with learning impacts, with being above the median for the sample being associated with 0.17 standard deviation higher impacts on learning.

### 4.3. Qualitative Analysis

We supplement the quantitative results with an analysis of self-reported perceptions by program implementers about the elements 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 up 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. Table 8 presents

the results for the program elements that implementers say work. Four of 13 interviewees – the most common response – mentioned that mentoring follow-up visits were a crucial component in making their training work, casting doubt on the effectiveness of face-to-face initial training in the absence of follow-up. The next most commonly perceived elements responsible for positive effects on student learning 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 – which were both mentioned by 3 of 13 interviewees. The other program element which received multiple mentions was the materials provided by the program, cited by 2 of 13 interviewees.

We asked similar questions of the program implementers about the program elements they perceived that teachers liked and disliked the most about their training programs, coding the responses in the same way and, interestingly, we only found one common response in each case. Tables 9 and 10 present those program elements which implementers reported were popular and unpopular among teachers, respectively. Three of the 13 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, discussion, and workshops rather than passive learning. Nine other program elements were mentioned by one implementer each as being popular among teachers, including the provision of concrete steps to follow, lesson plans, and materials, as well as a number of variables which are in a sense outcomes themselves, such as increased confidence and knowledge about content and technology.

Similarly, for unpopular program elements, 3 of the 13 program implementers we interviewed reported that teachers disliked the amount of time taken by participating in the training programs, which they perceived as excessive. In addition, one interviewee also mentioned that teachers complained about the high degree of control that the program exerts over how they exactly they use their time, rather than the amount of time committed to the program per se. Beyond this, another 6 program elements were mentioned once each, with the most widely applicable of these including that training was too theoretical and that training was not linked to the curriculum.

## 5. Conclusions

Governments spend enormous amounts of time and money on in-service teacher training. Many countries have many different in-service teacher training programs running simultaneously, and it is likely that many are ineffective. This paper demonstrates that there are associations with some characteristics of teacher training programs and student test score gains, such as the inclusion of supplemental materials, follow-up visits, and focus on a specific subject. However, this review reveals broad weakness in reporting on these 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 them often seeks to reduce them to a small handful of variables, and each scholar decides independently which variables are most relevant.

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 on-line. It is only through developing a more thorough and consistent understanding of teacher training programs that practitioners can learn what to emulate from the best programs and what to avoid from the worst.

## 6. References


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.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*: 1235–1265.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. (2009). Introduction to Meta-Analysis. Chichester, UK: John Wiley & Sons.

Bourgeois, E., & Nizet, J. (1997). Aprendizaje y Formación de Personas Adultas. París: Presses Universite de France.

Cardemil, C. (2001). Procesos y Condiciones en el Aprendizaje de Adultos. Santiago: MINEDUC.

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

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

Darling-Hammond, L. Wei, R. C. and Andree, A. (2010). How High-Achieving Countries Develop Great Teachers. Stanford Center for Opportunity Policy in Education Research Brief. August 2010.

Evans, D. K. & Popova, A. Forthcoming. What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews. *World Bank Research Observer*.

Fryer, R. G. (2016). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. NBER Working Paper No. 22130.

Gersten, R., Taylor, M. J., Keys, T.D., Rolfhus, E., and Newman-Gonchar, R. 2014. Summary of research on the effectiveness of math professional development approaches. (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

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

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

Knowles, M., Holton, E., & Swanson, R. (2005). The Adult Learner. Oxford: 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.

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

Kremer, M., Duflo, E. & Dupas, P. (2011). Peer effects, teacher incentives, and the impact of tracking. *American Economic Review, 101*: 1739 -1774. doi:10.1257/aer.101.5.1739

La Paro, K. M., & Pianta, R. C. (2003). CLASS: Classroom Assessment Scoring System. Charlottesville: University of Virginia.

McEwan, P. (2014). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research, 85*(3): 353-394.

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

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.

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

Stallings, J. (1977). Learning to Look: A Handbook on Classroom Observation and Teaching Models. Wadsworth Publishing: Belmont, CA.

TNTP (The New Teacher Project). (2015). The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development.

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

Wei, R.C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad. Dallas, TX. National Staff Development Council.

Woods, F., McQuarrie, F. (1999). On the job learning. *Journal of Staff Development*, *20*(3): 10-13.

World Bank. (2013). What matters most for teacher policies: A framework paper. SABER Working Paper Series. Number 4.

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, REL 2007–No. 033).

## Tables and Figures

**Figure 1: Search process and results**

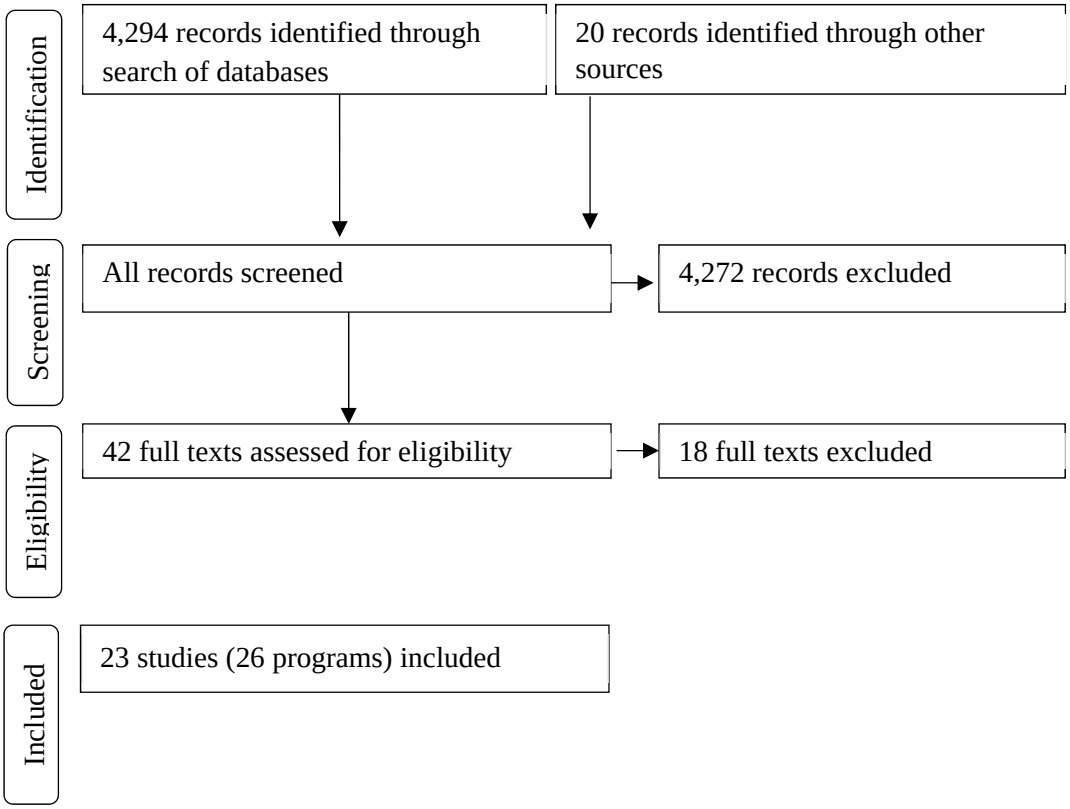


**Figure 2: Summary of the in-service teacher training survey instrument (ITTSI)**

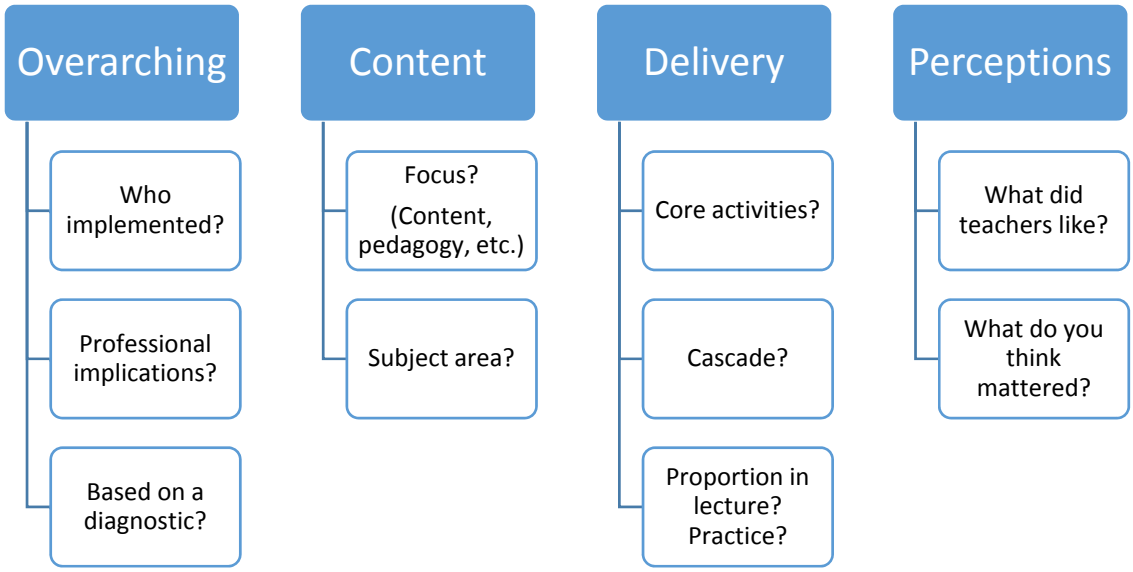

**Table 1. Data collection from studies vs. interviews**

| | Data collected | | | | |
|---|---|---|---|---|---|
| | From impact evaluation reports only | | After interviews with implementers | | |
| | Number of indicators | Percentage of indicators | Number of indicators | Percentage of indicators | Total indicators |
| Overarching aspects | 15.6 | 60% | 20.8 | 80% | 27 |
| Content | 11.3 | 43% | 17.4 | 67% | 10 |
| Delivery | 9.4 | 36% | 17.7 | 68% | 14 |
| TOTAL | 25.50 | 50% | 38 | 74% | 51 |
| For interviewed programs only: | | | 50 | 98% | 51 |

**Table 2. Overarching Aspects – Descriptive statistics**

| Overarching Aspects variable | Mean | Standard deviation |
|---|---|---|
| Teachers pay some cost for the training (including their own transport) = 0 | 0 | 0 |
| Number of teachers receiving training under this program in the last year | 609.0714 | 1619.068 |
| Number of schools in which the program was implemented | 56.6667 | 44.098 |
| Number of years the program been running | 2.9605 | 3.087 |
| Proportion of teachers who dropped out of the program in the last year | 0.1739 | 0.2774 |
| Targeting by geography | 0.5417 | 0.509 |
| Targeting by subject | 0.25 | 0.4423 |
| Targeting by grade | 0.7917 | 0.4149 |
| Targeting by years of experience | 0.0417 | 0.2041 |
| Targeting by skill gaps | 0 | 0 |
| Targeting by contract teachers | 0.125 | 0.3378 |
| Program provides materials | 0.8182 | 0.3948 |
| Program provides textbooks | 0.0952 | 0.3008 |
| Program provides storybooks | 0.2857 | 0.4629 |
| Program provides computers | 0.1429 | 0.3586 |
| Program provides teacher manuals | 0.5714 | 0.5071 |
| Program provides lesson plans/videos | 0.381 | 0.4976 |
| Program provides scripted lessons | 0.0952 | 0.3008 |
| Program provides craft materials | 0.1429 | 0.3586 |
| Program provides other reading materials (flashcards, word banks, primers) | 0.3333 | 0.483 |
| Program provides software | 0.2727 | 0.4558 |
| Program designed by Government | 0.1538 | 0.3679 |
| Program designed by NGO or social enterprise | 0.3846 | 0.4961 |
| Program designed by researchers | 0.4615 | 0.5084 |
| Program implemented by Government | 0.2692 | 0.4523 |
| Program implemented by NGO or social enterprise | 0.3846 | 0.4961 |
| Program implemented by researchers | 0.3462 | 0.4852 |
| Program design not based on any kind of diagnostics | 0.2353 | 0.4372 |
| Program design based on informal diagnostic | 0.3529 | 0.4926 |
| Program design based on formal diagnostic | 0.4118 | 0.5073 |
| Participation has no implications for salary or promotion | 0.4706 | 0.5145 |
| Participation has implications for status only | 0.1176 | 0.3321 |
| Participation has implications for promotion or points towards promotion or salary implications | 0.4118 | 0.5073 |
| Teachers are not evaluated | 0.3333 | 0.488 |
| Positive consequence if teachers are well evaluated | 0.2 | 0.414 |
| No positive consequence if teachers are well evaluated | 0.4667 | 0.5164 |
| Negative consequence if teachers are well evaluated | 0.1333 | 0.3519 |
| No negative consequence if teachers are well evaluated | 0.5333 | 0.5164 |

**Table 3. Content – Descriptive statistics**

| Content variable | Mean | Standard deviation |
|---|---|---|
| Primary focus is subject content | 0.0833 | .2823 |
| Primary focus is pedagogy | 0.4583 | .509 |
| Primary focus is technology | 0.2917 | .4643 |
| Primary focus is counseling | 0.0833 | .2823 |
| Primary focus is classroom management | 0.0417 | .2041 |
| Secondary focus is subject content | 0.6842 | .4776 |
| Secondary focus is pedagogy | 0.2105 | .4189 |
| Secondary focus is technology | 0.1053 | .3153 |
| No subject focus | 0.087 | .2881 |
| Subject focus is literacy/language | 0.6957 | .4705 |
| Subject focus is math | 0.2174 | .4217 |
| Subject focus is science | 0.0435 | .2085 |
| Subject focus is information technology | 0.0435 | .2085 |
| Training involves lectures | 0.9286 | .2673 |
| Training involves discussion | 0.6429 | .4972 |
| Training involves lesson enactment | 0.5714 | .5136 |
| Training involves materials development | 0.2857 | .4688 |
| Training involves training on how to conduct diagnostics | 0.3333 | .488 |
| Training involves lesson planning | 0.5789 | .5073 |
| Training involves the use of scripted lessons | 0.3333 | .4851 |

**Table 4. Delivery – Descriptive statistics**

| Delivery variable | Mean | Standard deviation |
|---|---|---|
| Total hours of face-to-face training | 64.04 | 45.5764 |
| Over how many weeks? | 11.79 | 14.5273 |
| Over how many months? | 11.55 | 8.2874 |
| Proportion of training spent in lectures | 0.478 | .2917 |
| Proportion of training spent practicing with other teachers | 0.515 | .3192 |
| Proportion of training spent practicing with students | 0.062 | .0829 |
| How many in-school follow-up support visits do teachers receive after the initia | 5.833 | 9.1684 |
| How many weeks of distance learning does the program include (if any)? | 1.438 | 4.3354 |
| Is it a cascade training model (i.e. one where program trainers trainers who the | 0.5 | .513 |
| Is there a part of the training where teachers meet with trainers for several da | 0.92 | .2769 |
| Includes follow up visits | 0.778 | .4278 |
| Includes distance learning | 0.188 | .4031 |
| Trainers are primary or secondary teachers | 0.222 | .4278 |
| Trainers are experts - university professors or have graduate degrees in educati | 0.333 | .4851 |
| Trainers are researchers | 0.056 | .2357 |
| Trainers are local government officials | 0.333 | .4851 |
| Training held at schools | 0.063 | .25 |
| Training held at central location including hotel conference room etc. | 0.813 | .4031 |
| Training held at university or training center | 0.125 | .3416 |
| Follow-up visits for in-class pedagogical support | 0.412 | .5073 |
| Follow-up visits for monitoring | 0.471 | .5145 |
| Follow-up visits to review material | 0.118 | .3321 |

**Table 5. Overarching Aspects – 5 largest regression coefficients**

| Overarching Aspects variable | Program impact on student learning |
|---|---|
| Program provides textbooks | 0.355** |
| | (0.128) |
| Program provides other reading materials (flashcards, word banks, primers) | 0.159* |
| | (0.087) |
| Participation has implications for promotion or points towards promotion or salary implications | 0.143** |
| | (0.066) |
| Targeting by years of experience | 0.136 |
| | (0.198) |
| Program provides storybooks | 0.129 |
| | (0.094) |
| Observations | 26 |

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

**Table 6. Content – 5 largest regression coefficients**

| Content variable | Program impact on student learning |
|---|---|
| Primary focus of the training program is classroom management | 0.471 |
| | (0.272) |
| No subject focus of training | -0.243 |
| | (0.204) |
| Secondary focus of the training program is subject content | 0.182 |
| | (0.156) |
| Primary focus of the training program is new technology | 0.180 |
| | (0.206) |
| Primary focus of the training program is pedagogy | 0.177 |
| | (0.201) |
| Observations | 26 |

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

**Table 7. Delivery – 5 largest regression coefficients**

| Delivery variable | Program impact on student learning |
|---|---|
| Training takes place in university or training center | 0.385** |
| | (0.142) |
| Follow-up visits to review material | 0.256 |
| | (0.156) |
| Most common profile of the direct trainers is researchers | -0.196 |
| | (0.336) |
| Most common profile of the direct trainers is local government officials | -0.170 |
| | (0.257) |
| Proportion of training spent practicing with other teachers | 0.169 |
| | (0.134) |
| Observations | 26 |

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

**Table 8. Trainers' perceptions of most effective program elements**

| Program element | Tally |
|---|---|
| Mentoring follow-up visits | 4 |
| Engaging teachers for their opinions and ideas - either through discussion or text messages | 3 |
| Programs designed in response to local context - building on what teachers already do & linking to everyday experiences | 3 |
| Materials | 2 |
| Assessment | 1 |
| Giving status to teachers | 1 |
| Making learning fun for teachers and students | 1 |
| Quality of training | 1 |
| Support of headmasters and mayor | 1 |
| Cascade model | 1 |
| Making teachers aware that they should communicate with students | 1 |
| Establishing agreed upon characteristics for what makes a good lesson | 1 |
| Time between sessions for teachers to work independently | 1 |
| Use of open source math software | 1 |

Tally reflects the number of program implementers (of the 13 interviewed in total) who mentioned in an open-ended question that they believed a given program element was one of the most important for the effectiveness of their in-service teacher training program.

**Table 9. Popular program elements among teachers**

| Program element | Tally |
|---|---|
| Fun - including participation, discussion, and workshops rather than traditional lecture format with passive learning | 3 |
| Concrete steps to follow | 1 |
| Improved their content knowledge | 1 |
| Increased group work | 1 |
| Methodology | 1 |
| Lesson plans | 1 |
| Learning to use computers ahead of the curve | 1 |
| Motivational video | 1 |
| Gave confidence to teachers | 1 |
| Materials | 1 |

Tally reflects the number of program implementers (of the 13 interviewed in total) who mentioned in an open-ended question that they believed a given program element was particularly liked by teachers.

**Table 10. Unpopular program elements among teachers**

| Program element | Tally |
|---|---|
| High demands on their time | 3 |
| Courses did not cover some newer material | 1 |
| Giving children information on the costs of schooling | 1 |
| High control of how they use their time | 1 |
| Not all schools and years received the program | 1 |
| Teaching in English | 1 |
| Too theoretical | 1 |
| Training not linked to curriculum | 1 |

Tally reflects the number of program implementers (of the 13 interviewed in total) who mentioned in an open-ended question that they believed a given program element was particularly disliked by teachers.

## Appendix A: The List of Included Papers


1. Abeberese, A. B., T. J. Kumler, & Linden, L. L. (2011). "Improving Reading Skills by Encouraging hildren to Read in School: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines." National Bureau of Economic Research (NBER Working Paper No. 17185).
2. 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.
3. Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, & 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.
4. Banerjee, A. V., S. Cole, E. Duflo, & Linden, L. L. (2007). "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics, 122*(3): 1235-1264.
5. 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 Series 4836.
6. Beuermann, D., E. Naslund-Hadley, I. J. Ruprah, & Thompson, J. (2012). "The Pedagogy of Science and Environment: Experimental Evidence from Peru." Inter-American Development Bank Working Paper OVE/WP-01/12.
7. Brooker, S., H. Inyega, B. Estambale, K. Njagi, E. Juma, C. Jones, C. Goodman, & Jukes, M. (2013). Impact of malaria control and enhanced literacy instruction on educational outcomes among Kenyan school children: a multi-sectoral, prospective, randomised evaluation, 3ie Draft Grantee Final Report, International Initiative for Impact Evaluation (3ie).
8. Carrillo, P., M. Onofa, & Ponce, J. (2010). "Information Technology and Student Achievement: Evidence from a Randomized Experiment in Ecuador." IDB Working Paper Series No. IDB-WP-223.
9. He, F., L. L. Linden, & MacLeod, M. (2008). How to Teach English in India: Testing the Relative Productivity of Instruction Methods within the Pratham English Language Education Program.
10. Kerwin, J. T., & Thornton, R. (2015). Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda.
11. Lai, F., R. Luo, L. Zhang, X. Huang, & Rozelle, S. (2011). "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing." Rural Education Action Project Working Paper 228.
12. Lai, F., L. Zhang, Q. Qu, & Hu, X. (2012). "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Public Schools in Rural Minority Areas in Qinghai, China."

13. Loyalka, P., C. Liu, Y. Song, H. Yi, X. Huang, J. Wei, L. Zhang, Y. Shi, J. Chu, & Rozelle, S. (2013). "Can information and counseling help students from poor rural areas go to high school? Evidence from China." *Journal of Comparative Economics, 41*(4): 1012-1025.

14. Lucas, A. M., P. J. McEwan, M. Ngware, & Oketch, M. (2014). "Improving Early-grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management, 33*(4): 950-976.

15. Mo, D., L. Zhang, R. Luo, Q. Qu, W. Huang, J. Wang, Y. Qiao, M. Boswell, & Rozelle, S. (2014). "Integrating computer-assisted learning into a regular curriculum: evidence from a randomised experiment in rural schools in Shaanxi." *Journal of Development Effectiveness, 6*(3): 300-323.

16. Nitsaisook, M., & Anderson, L. W. (1989). "An Experimental Investigation of the Effectiveness of Inservice Teacher Education in Thailand." *Teaching and Teacher Education, 5*(4): 287-302.

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

18. Pournara, C., J. Hodgen, J. Adler, & Pillay, V. (2015). "Can improving teachers' knowledge of mathematics lead to gains in learners' attainment in Mathematics?" *South African Journal of Education, 35*(3).

19. 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. Washington, DC, USAID.

20. Tan, J.-P., J. Lane, & Lassibille, G. (1999). "Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments." *The World Bank Economic Review, 13*(3): 493-508.

21. Weir, C. J. , & Roberts, J. (1991). Evaluating a Teacher Training Project in Difficult Circumstances. Issues in Language Programme Evaluation in the 1990's. S. Anivan.

22. 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.

23. Zhang, L., F. Lai, X. Pang, H. Yi, & 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.

## Appendix B: The Survey Instrument

Introduction

Thank you for participating in our survey. Your feedback is important.

**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?**

◯ National ◯ Multiple states or regions ◯ One state or region ◯ Less than one state or region

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

☐ Government
☐ Non-governmental organization
☐ Private company or social enterprise
☐ Researchers

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

☐ Government
☐ Non-governmental organization
☐ Private company or social enterprise
☐ 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?**

◯ Teacher training is primary focus ◯ 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?**

◯ No ◯ Yes, informal diagnostic ◯ Yes, formal diagnostic

**12. Was the program design based on a diagnostic or evaluation of teacher skills of some kind? If so, what kind?**

◯ No ◯ Yes, informal diagnostic ◯ Yes, formal diagnostic

**13. What teacher skill gaps is this program designed to support?**

◯ Subject content ◯ Subject-specific pedagogy ◯ Technology ◯ Counseling ◯ Classroom management

◯ Specific tool ◯ Assessment ◯ Curricular update ◯ General pedagogy ◯ Theory

**14. Is the program for all teachers or just for certain teachers?**

◯ All teachers ◯ Certain teachers

**15. If the program is just for certain teachers, on what characteristics is it targeted? (Select all that apply.)**

☐ Geography ☐ Subject ☐ Grade ☐ Teachers' years of experience ☐ Teachers' skill gaps ☐ Uncertified teachers

☐ Contract teachers

**16. Which grades?**

☐ Grade 1 ☐ Grade 2 ☐ Grade 3 ☐ Grade 4

☐ Grade 5

**17. Are teachers assigned to participate or do they volunteer for the program?**

☐ Assigned ☐ Volunteer ☐ A mix of both

**18. How much do teachers have to pay to register for the program (if anything) per year, in U.S. dollars?**

**19. Which of the following other costs do the teachers have to pay to participate in the program? (Select all that apply.)**

☐ None ☐ Transport ☐ Accommodation ☐ Materials

☐ Other

**20. How much do teachers receive as per diem or payment to participate in the program per year, in U.S. dollars?**

**21. What is the total cost of the program per year, in U.S. dollars?**

**22. Does participation in the training program have any professional implications for teachers? (Select all that apply.)**

- [ ] No
- [ ] Status
- [ ] Promotion or points towards promotion
- [ ] Salary
- [ ] Official certification

**23. Are teachers evaluated at the end of the training?**

- ( ) No
- ( ) Yes

**24. Is it possible for teachers to fail this exam?**

- ( ) No
- ( ) 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.)**

- [ ] No
- [ ] Status
- [ ] Promotion or points towards promotion
- [ ] Salary
- [ ] Official certification

**27. Is there a negative consequence if teachers are poorly evaluated? (Select all that apply.)**

- [ ] No
- [ ] Status
- [ ] Promotion or points towards promotion
- [ ] Salary
- [ ] Official certification

**28. Which of the following are informed about the teachers' performance on the training evaluation? (Select all that apply.)**

- [ ] None
- [ ] Teacher
- [ ] School where the teacher teaches
- [ ] Ministry of Education

**29. What materials, if any, did the program provide alongside the training? (Select all that apply.)**

- [ ] No materials
- [ ] Word banks
- [ ] Lesson plans/videos
- [ ] Textbooks
- [ ] Computers
- [ ] Scripted lessons
- [ ] Storybooks or reading pamphlets
- [ ] Software
- [ ] Craft materials
- [ ] Flashcards
- [ ] Teacher manuals

**30. How many teachers are receiving training under this program this year?**

**31. In the last year, 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?**

- ◯ No
- ◯ Yes

**34. If so, on which of the following was it evaluated in terms of impact? (Select all that apply.)**

- [ ] Teacher knowledge
- [ ] Teacher behavior
- [ ] Student learning
- [ ] Objectives of the program

**35. Over the course of the program, what data are collected centrally?**

- [ ] Frequency of class delivery
- [ ] Attendance of participating teachers
- [ ] Teachers' assessment of value of training
- [ ] Test score of teacher subject knowledge
- [ ] Test score of teacher pedagogical knowledge
- [ ] Practical test observing teaching

## Content

**36. Which of these is the primary focus of the training program?**

◯ Subject content ◯ Subject-specific pedagogy ◯ Technology ◯ Counseling ◯ Classroom management

◯ Specific tool ◯ Assessment ◯ Curricular update ◯ General pedagogy ◯ Theory

**37. Which of these is the secondary focus of the training program?**

◯ No other focus ◯ Subject content ◯ Subject-specific pedagogy ◯ Technology ◯ Counseling

◯ Classroom management ◯ Specific tool ◯ Assessment ◯ Curricular update ◯ General pedagogy

◯ Theory

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

☐ None ☐ Literacy or language ☐ Math ☐ Natural science ☐ Social science ☐ Information technology

☐ Other

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

◯ In-person ◯ Online ◯ 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?**

◯ No ◯ Yes

**45. If so, is the contact with trainers individual, in groups, or both?**

◯ Individual ◯ Group ◯ Both

**46. Are the online group sessions compulsory or voluntary?**

◯ Compulsory ◯ 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.)**

- [ ] Lectures
- [ ] Discussion
- [ ] Teaching practice
- [ ] Discussion of videos
- [ ] Practice in science labs
- [ ] Practice with computers
- [ ] Other practical activities

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

- [ ] None
- [ ] Development of pedagogical materials
- [ ] Development of classroom evaluation materials
- [ ] Training on how to conduct diagnostics
- [ ] Lesson planning
- [ ] Using scripted lessons

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

- ◯ No
- ◯ Yes

**51. What is the most common profile of the trainers or facilitators who the teachers have direct contact with?**

- ◯ Primary or secondary teacher in the subject of the training
- ◯ Specially selected expert primary or secondary teacher
- ◯ Other primary or secondary teacher
- ◯ University professor or graduate degree in education
- ◯ Researcher
- ◯ Government official
- ◯ University student in education
- ◯ 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.)**

- [ ] None
- [ ] Designed the program
- [ ] Received a specific certification
- [ ] Received one week or less of training
- [ ] 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.)**

- [ ] None
- [ ] Performance related bonus
- [ ] Tablet or computer
- [ ] Books
- [ ] Community recognition
- [ ] 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 the these types of follow-up support do teachers receive? (Select all that apply.)**

- [ ] Text messages
- [ ] Phone calls
- [ ] Emails
- [ ] In-school support from principals
- [ ] 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?**

- ( ) No
- ( ) 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?**

- ◯ School of teacher being trained
- ◯ Central location (other school, hotel, government building etc.)
- ◯ 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.)**

- [ ] In-class pedagogical support
- [ ] Monitoring
- [ ] 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.)**

## Perceptions

**71. Were there any elements of the program that the teachers particularly *liked*?**

| | |
|---|---|
| Element 1 | |
| Element 2 | |
| Element 3 | |

**72. Were there any elements of the program that the teachers particularly *disliked*?**

| | |
|---|---|
| Element 1 | |
| Element 2 | |
| Element 3 | |

**73. What were the key elements you think made the program work?**

| | |
|---|---|
| Element 1 | |
| Element 2 | |
| Element 3 | |

## 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, $S_{pooled}$. While all studies report $D$ directly, $S_{pooled}$ 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:

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

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:

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

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

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

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

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

## Appendix D: Complete Regression Results

**Table C1. Overarching Aspects – complete regression results**

| Overarching Aspects variable | Program impact on student learning |
|---|---|
| Program provides textbooks | 0.355** |
| | (0.128) |
| Program provides other reading materials (flashcards, word banks, primers) | 0.159* |
| | (0.087) |
| Participation has implications for promotion or points towards promotion or salary implications | 0.143** |
| | (0.066) |
| Targeting by years of experience | 0.136 |
| | (0.198) |
| Program provides storybooks | 0.129 |
| | (0.094) |
| Number of schools in which the program was implemented | 0.120 |
| | (0.083) |
| Program designed by Government | 0.100 |
| | (0.130) |
| Participation has implications for status only | 0.095 |
| | (0.101) |
| Teachers dropped out of program in last year | 0.088 |
| | (0.096) |
| Teachers not evaluated | -0.084 |
| | (0.084) |
| Program provides craft materials | -0.084 |
| | (0.126) |
| Positive consequence if teachers are well evaluated | 0.080 |
| | (0.100) |
| Program provides materials | 0.078 |
| | (0.109) |
| Program designed by NGO or social enterprise | 0.065 |
| | (0.086) |
| Number of teachers receiving training under this program in the last year | 0.065 |
| | (0.088) |
| Number of years the program been running | 0.063 |
| | (0.119) |
| Negative consequence if teachers are poorly evaluated | 0.054 |
| | (0.117) |
| Targeting by grade | -0.050 |
| | (0.098) |

| Overarching Aspects variable | Program impact on student learning |
|---|---|
| Program implemented by NGO or social enterprise | 0.039 |
| | (0.094) |
| Program provides computers | 0.035 |
| | (0.127) |
| Targeting by subject | -0.034 |
| | (0.099) |
| Program provides teacher manuals | -0.034 |
| | (0.090) |
| Targeting by contract teachers | 0.027 |
| | (0.121) |
| Program provides scripted lessons | 0.020 |
| | (0.151) |
| Program implemented by Government | -0.017 |
| | (0.107) |
| Program provides lesson plans/videos | -0.017 |
| | (0.091) |
| Program design based on informal diagnostic | -0.012 |
| | (0.134) |
| Program provides software | 0.011 |
| | (0.096) |
| Program design not based on any kind of diagnostics | 0.006 |
| | (0.151) |
| Targeting by geography | -0.001 |
| | (0.082) |
| Observations | 26 |
| Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. | |

**Table C2. Content – complete regression results**

| Content variable | Program impact on student learning |
|---|---|
| Primary focus of the training program is classroom management | 0.471 |
| | (0.272) |
| No subject focus of training | -0.243 |
| | (0.204) |
| Secondary focus of the training program is subject content | 0.182 |
| | (0.156) |
| Primary focus of the training program is new technology | 0.180 |
| | (0.206) |
| Primary focus of the training program is pedagogy | 0.177 |
| | (0.201) |
| Secondary focus of the training program is pedagogy | 0.123 |
| | (0.178) |
| Training involves lesson enactment | 0.110 |
| | (0.072) |
| Training involves lesson planning | 0.096 |
| | (0.102) |
| Primary focus of the training program is counseling | -0.089 |
| | (0.235) |
| Training involves training on how to conduct diagnostic | 0.083 |
| | (0.074) |
| Primary focus of the training program is subject content | 0.083 |
| | (0.235) |
| Subject focus of the training program is science | -0.063 |
| | (0.250) |
| Subject focus of the training program is information technology | 0.062 |
| | (0.250) |
| Training involves the use of scripted lessons | 0.052 |
| | (0.108) |
| Subject focus on literacy or language | 0.029 |
| | (0.155) |
| Subject focus of the training program is math | -0.028 |
| | (0.186) |
| Training involves lectures | 0.020 |
| | (0.152) |
| Training involves materials development | 0.012 |
| | (0.087) |
| Training involves discussion | 0.004 |
| | (0.082) |

| | |
|---|---|
| Observations | 26 |
| Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. | |

**Table C3. Delivery – complete regression results**

| Delivery variable | Program impact on student learning |
|---|---|
| Training takes place in university or training center | 0.385** |
| | (0.142) |
| Follow-up visits to review material | 0.256 |
| | (0.156) |
| Most common profile of the direct trainers is researchers | -0.196 |
| | (0.336) |
| Most common profile of the direct trainers is local government officials | -0.170 |
| | (0.257) |
| Proportion of training spent practicing with other teachers | 0.169 |
| | (0.134) |
| Includes consecutive days of face-to-face training | 0.165 |
| | (0.148) |
| Proportion of training spent in lectures | -0.156 |
| | (0.094) |
| Includes follow-up visits | 0.146 |
| | (0.122) |
| Most common profile of the direct trainers is university professors or graduate degrees in education | -0.146 |
| | (0.257) |
| Follow-up visits for in-class pedagogical support | 0.144 |
| | (0.102) |
| Months over which distance learning is spread | -0.140 |
| | (0.085) |
| Most common profile of the direct trainers is primary or secondary teachers | -0.094 |
| | (0.274) |
| Includes distance learning | -0.066 |
| | (0.092) |
| Total hours of face-to-face training | 0.058 |
| | (0.083) |
| Cascade training model | -0.039 |
| | (0.100) |
| Proportion of training spent practicing with students | 0.033 |
| | (0.093) |
| Weeks over which face-to-face training is spread | -0.025 |
| | (0.085) |
| Training takes place in-school | 0.004 |

| | |
|---|---|
| | (0.194) |
| Observations | 26 |
| Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. | |