Why Programs Fail: Lessons for Improving Public Service Quality from a Mixed-Methods Evaluation of an Unsuccessful Teacher Training Program in Nepal

Julie Schaffner
The Fletcher School, Tufts University

Paul Glewwe
Department of Applied Economics, University of Minnesota

Uttam Sharma
Independent Consultant

November 12, 2021

Abstract: Using a randomized control trial embedded within a mixed-methods evaluation, we find that an at-scale government teacher training program, of a common but seldom-evaluated form, has little or no impact on student learning. We then document five challenges that the policy's design failed to address, related to: oversight of training sessions, school-level difficulties in releasing teachers for training (lack of substitute teachers), deficits in teachers' subject knowledge, deficits in teachers' post-training accountability and support, and students' needs for differentiated instruction. We discuss implications for the literatures on teacher training program design and on good governance for public service provision.

Acknowledgements: We gratefully acknowledge funding from the International Initiative for Impact Evaluation (3ie) under Grant PW3.10.NP.IE. We thank collaborators in the Government of Nepal's National Planning Commission, especially then-Joint Secretary Teertha Dhakal, and in the Ministry of Education, Science and Technology, and the Center for Education and Human Resource Development. We are grateful for research and logistical assistance provided by Deepika Shrestha, Sunil Poudel, Tri Bikram Pandey, and Tejkala Uprety of the Center for Policy Research and Consultancy; for research assistance from Girija Bahety, Floor de Ruijter, Kathryn Hirschboeck, Chhavi Kotwani, Rayyan Mobarak, Silver Namunane, and Jeremy Schlitz; for qualitative research guidance from Sushan Acharya; and for student assessment development by Aditi Bhomick, Alejandro Ganimian, Peshal Khanal, Krishna Prasad Adhikari, and Kamal Prasad Acharya.

Policymakers around the world invest heavily in teacher training, hoping to raise teaching quality, which is critical for successful and inclusive student learning (Rivkin, Hanushek and Kain 2005; Chetty, Friedman and Rockoff 2014; Araujo et al. 2016).[1] Evaluations of some teacher training programs in low- and middle-income countries document substantial learning impacts, indicating that training program success is possible (Albornez, et al. 2020, Beg, et al. 2019, Cilliers, et al. 2020a). Evaluations of other training programs, however, show little impact and even reduced learning for some students, indicating that success is far from guaranteed (Loyalka, et al. 2019 and Blimpo and Pugatch 2021). Rigorous evaluations such as these remain scarce, however, and much remains to be learned about how to design and implement successful teacher training programs.[2] Improving educational outcomes in these countries will require not only more estimates of the impacts of different types of teacher training programs, but also a better understanding of the obstacles that may inhibit training program success and identification of features of program design and governance that policymakers may use to overcome those obstacles.

Scrutiny of teacher training programs is important in Nepal, where student learning in government schools remains low. Learning deficits are especially pronounced among secondary students in government schools. In results presented below, for example, we find that nearly 40% of 9th grade students in government schools are unable to calculate the area of a square (with sides 3 cm. in length), over 30% are unable to solve a simple story problem involving single-digit multiplication, and over 20% cannot find the solution to the third-grade problem of 6 ÷ 3. The 2019 Secondary Education Examination (SEE), the nationwide school leaving exam administered at the end of 10th grade, documents low learning levels more broadly, especially among students in government schools; only 4.3% of government school students scored at the "Very good"/B+ level or higher, while 16% scored below the "Acceptable"/C level. (Dixit 2019).[3]

We study the impacts of a training program for teachers of 9th and 10th grade math and science in Nepal's government schools. The government rolled out the trainings between December 2017 and May 2018, using governance arrangements that were standard for at-scale government teacher trainings over the preceding decade. The trainings sought to improve teachers' understanding of challenging math or science concepts in the 9th and 10th grade curricula, and to encourage teachers to use methods for teaching these concepts that involve demonstrations, often with teaching aids made from local materials. The program required teachers to attend 10 days of in-person training at Education Training Centers (ETCs). It also required them to complete five days of self-study project work, including the creation of 10 lesson plans

[1] Loyalka, et al. (2019) report that China spent over $1 billion per year on in-service teacher training, that India spent $1.2 billion on such training between 2012 and 2017, and that the average teacher in Mexico spends 23 days per year in teacher training. Between 2000 and 2010 "nearly two thirds" of World Bank-supported education projects included teacher training (Popova, et al. 2019).

[2] As discussed in Part V, we found only 23 high-quality evaluations of relevant teacher training programs.

[3] It is unlikely that the scores are low because of an unrealistically high grading standard, given that 41.4% of private school students (who constitute 15.2% of exam takers) scored at the "Very good"/B+ level or higher (Republica 2019).

and related teaching aids, after returning to their schools. These trainings are important components of Nepal's School Sector Development Program (SSDP), the government's overall plan to raise school quality and inclusivity between 2016 and 2023, and such teacher trainings are common around the world.[4]

We employ a cluster-randomized experimental design to estimate the impacts of the SSDP trainings on math and science test scores for 9th and 10th grade government school students, using a (nearly) nationally representative, stratified sample of 203 schools in 16 of Nepal's 75 districts, with data on nearly 10,000 students. Our sample is drawn from the population of schools that include both basic (1 to 8) and secondary (9 and 10) grades, in which 97% of Nepal's 9th and 10th grade government school students receive their education.[5] Schools were randomly allocated to treatment and control within strata defined by district and schools' pre-program teacher training levels, and treatment and control schools are well balanced. We estimate impacts on eight Item Response Theory (IRT)-based student test score variables, distinguishing between: two grades (9 and 10), two subjects (math and science), and two types of score for the assessments: (1) a total score based on answers to all 35 questions in the assessment, and (2) an "SSDP focus" score based on answers to the subset of questions that we deemed most closely tied to the SSDP training curricula. In most cases, the trainings took place early in the 2018-19 school year, and tests were administered at the end of that year. We present ITT estimates using OLS regressions on the full endline sample without baseline test score controls and on a smaller panel sample controlling for baseline test scores. All these regressions employ population weights and include controls for test-taking conditions, as explained below.

We demonstrate, first, that the training programs had at most only small positive effects on student test scores. Of our eight impact estimates in the full endline sample, five of the point estimates are negative (one of them statistically significantly so at the 10% level), and 95% confidence intervals rule out average impacts above 0.10 standard deviations (of the control groups' test score distribution) in three of eight cases, and above 0.18 standard deviations in all cases. While attrition rates are high, near 40%, they differ little between the treatment and control samples. The results are robust to use of unweighted rather than weighted regressions, omission of test-taking condition indicators, use of normalized raw percentage test scores rather than IRT-based indices of assessment performance, and inclusion of baseline controls. Standard errors are clustered at the school level, and results are robust to adjustment for multiple hypothesis testing.

Drawing on a rich mix of quantitative and qualitative evidence, we then examine the assumptions underlying the program's theory of change. We find evidence of five key weaknesses in program design and implementation that may explain the program's weak performance. First, from quantitative and qualitative evidence acquired through telephone interviews with trainers and training participants, we

[4] Popova, et al. (2019) report that at-scale teacher training interventions typically involve centralized trainings (73%) without follow-up in-class pedagogical support (63%) and without distribution of scripted lesson plans (67%).
[5] These numbers are based on Education Management Information System (EMIS) data from Nepal.

identify weaknesses in the governance of the ETC training sessions. While weak accountability and motivation among trainers (who expressed low expectations of program impact on training participants) may have diminished the quality of the trainings, we conclude that inadequate central-level efforts to ensure that trainers have the necessary skills and resources (including information, time, facilities, and materials) were at least as important in limiting training session quality. In particular, several of the 13 ETCs for which we have data used curriculum guidelines from previous trainings because they were unaware that new guidelines had been developed (or did not receive them), trainers reported having little time or guidance for preparing the training content , and some ETCs lacked trainers with adequate math and science expertise.

Second, rates of SSDP training participation among teachers in treated schools at endline were disappointingly low. Administrative data indicate that only 60% of math teachers and 42% of science teachers in the treatment schools at endline (and thus during the academic year most relevant for endline student learning outcomes) participated in the trainings. Participation rates were low in part because moderately high population rates of teacher turnover (19% for math teachers and 27% for science teachers between baseline and endline) meant that some teachers had joined the schools after the training had occurred, yet the most important reason for low participation is that a large share of the teachers in the schools at the time of the training did not take up the invitation to training. Qualitative evidence indicates that some schools refused to send their teachers to the trainings because the trainings took place on regular school days and substitute teachers for secondary math and science classes are difficult to find. Teachers' low expectations regarding the novelty and value of the trainings may have further reduced participation.

Third, we find evidence of serious weaknesses in some teachers' pre-training subject knowledge, which may have limited their ability to fully benefit from the SSDP training content focused on advanced math and science concepts. We indirectly tested teacher subject knowledge by asking teachers to evaluate anonymously a subsample of the questions that appeared on the student assessments. For each assessment item, the teachers were asked to select the response they believed the question's designer intended as the correct response and were also asked to evaluate the question's clarity and appropriateness to the relevant curriculum. One fifth of math teachers offered incorrect answers to a simple algebra problem, which raises questions about their preparation for SSDP training content related to more advanced algebra techniques, such as factoring polynomials. More concerning, nearly 40% of math teachers did not correctly calculate a rectangle's perimeter, suggesting an insufficient foundation for SSDP training content on surface areas of three-dimensional shapes. One fifth of science teachers incorrectly identified the main function of red blood cells, suggesting inadequate preparation for SSDP curriculum on the circulatory system, and nearly half of science teachers failed a question about evaporation, which is likely an important building block for SSDP content on climate change. We find no evidence that the SSDP trainings raised teachers' subject knowledge.

Fourth, while teachers reported finding at least some of the training content useful and expressed interest in implementing the new demonstration-based teaching methods, we find that the combination of daunting time requirements for integrating SSDP teaching practices into lessons throughout the school year, lack of teaching supplies, and lack of accountability for teaching innovation within schools left teachers insufficiently motivated to adopt the new teaching practices. Trainers expected that time constraints would prevent teachers from adopting the new practices, and teachers reported lack of time as an obstacle to adoption. Very few teachers reported being asked to report to anyone on their adoption of new teaching practices after the training. Most head teachers (school principals) have heavy teaching loads and seldom observe teachers in their classrooms. Even more telling, at endline almost none of the trained teachers could name even one of several self-study projects of which they were required to choose and complete one. This suggests that few teachers completed this self-study requirement. Institutions that fail to hold teachers accountable for completing these small projects are unlikely to do so for the larger investment needed to adopt new teaching practices. Teachers also reported difficulties in adopting new practices due to inadequate budgets for teaching materials, and difficulty in implementing new demonstration-based teaching methods in class periods that are typically only 40 to 45 minutes.

Fifth, we document that many students enter grades 9 and 10 with below-grade-level math and science skills. The SSDP trainings, which focused on new methods to teach advanced 9th and 10th grade math and science concepts, may, therefore, have equipped teachers with skills that are largely irrelevant to many students' learning needs. Working with both local and international expert developers of academic assessments, we incorporated questions appropriate for lower grade levels into our endline assessments. We find, for example, that over half of 9th graders at endline failed to correctly answer a question that local experts associate with the grade 8 mathematics curriculum in Nepal, suggesting that students had not fully learned or retained what they should have from grade 8. More worrying, over one third failed to solve a simple story problem requiring only single-digit multiplication, and over one fifth failed to correctly answer the question "What is 6 divided by 3?" which tests basic division at approximately the grade 3 level. Similar weaknesses appeared in their science preparation.

Finally, we offer rough estimates of program costs. At about $130 per teacher, or $2.60 to $3.00 per student, the SSDP trainings' cost is similar to that of interventions that have been found to raise student learning significantly in other contexts. See, for example, evidence on teacher incentive and student incentive programs reviewed in Damon et al. (2019). This suggests the importance for Nepal's policymakers of improving teacher trainings or replacing them with more cost-effective interventions.

This study contributes in three ways to the growing literature on the design and impacts of teacher training programs in low- and middle-income countries. First, we add to the small number of rigorous impact evaluations for government teacher training programs of a common form, which brings groups of

teachers to centralized training venues; instructs them in pedagogical theories or approaches but provides them with no lesson plans, teaching materials or other structured curriculum support; and includes no subsequent monitoring or coaching. Only three such programs have previously been the focus of rigorous impact evaluations. None of these studies finds a significantly positive impact on student learning, and two of the three studies rule out, with 95% confidence, anything more than small positive impacts. By adding a fourth null result to this list, our study strengthens the growing consensus regarding the need to re-design this prevalent form of government teacher training.

Second, our qualitative evidence offers additional support for, and new hypotheses regarding, two emerging themes in the literature, which highlight the importance for training program effectiveness of two design features: providing teachers with lesson plans and related teaching materials (or other structured curriculum support) that help teachers integrate new teaching approaches into all topics throughout the school year's curriculum; and the importance of following up initial group trainings with individual monitoring, coaching or other longer-term contact between teachers and training program personnel. Where teachers are time constrained, as our evidence suggests for Nepal, provision of lesson plans and other structured curriculum support may encourage adoption of new teaching practices, even among teachers who already have adequate skills, by reducing the substantial daily time costs associated with this adoption. At the same time, where neither schools nor district education institutions hold teachers accountable for good teaching practices, longer-term monitoring and coaching by training program personnel may, beyond any impact on teacher skills, improve teacher motivation by creating social or psychological rewards for teachers who adopt improved practices.

Third, we demonstrate how a mixed methods evaluation that examines the assumptions underlying a program's theory of change can reveal important practical challenges that policymakers will face when designing the content, delivery mode and governance of future teacher training programs. For economist "plumbers," as described by Duflo (2017), such systematic efforts to notice the details that may matter for policy impact can generate important hypotheses to test in future experiments. For Nepal, we highlight the challenges policymakers face of: providing good governance for trainers and training sessions; reducing the costs to schools of sending teachers to trainings (e.g. by scheduling trainings outside of regular school hours); devising ways to differentiate teachers by level of subject knowledge in order to *teach teachers* at the right level; reducing the costs of, and/or increasing the rewards for, teachers' adoption of new daily teaching practices; and devising training content that helps teachers *teach students* at the right level.

We also contribute to the larger literature on efforts to improve education quality, and to improve the governance of public service provision more generally. Since at least the publication of the *World Development Report 2004: Making Services Work for the Poor* (World Bank 2003), improving the "governance" of public service provision has often been conceptualized narrowly as a matter of improving

motivation by strengthening accountability relationships. More recent studies evaluating efforts to improve education quality, as reviewed in Glewwe and Muralidharan (2016), Snilsveit (2015) and World Bank (2018), are still largely shaped by this accountability-focused view of good governance. An independent literature on "management quality" similarly emphasizes accountability, highlighting the importance of good goal setting, performance monitoring and personnel management within schools and other frontline service facilities (Lemos and Scur 2016; Lemos, Muralidharan and Scur 2021).

Our findings, while consistent with the importance of ensuring accountability and motivation among public service providers, suggest an important way in which the current conceptualization of good governance for public services should be expanded. Specifically, good governance involves ensuring not only that agents have adequate *motivation*, but also that they have adequate *time*, *capacity*, *information*, and *budgets* for their work (Schaffner, 2014, Chapter 13). The practical implication is that, in addition to setting clear goals, monitoring performance, and tying rewards to good performance, policymakers and managers must adequately evaluate their agents' initial time, skill, information, and budget constraints. When defining agents' responsibilities, they must then take these constraints into account and either define responsibilities that are feasible given those constraints or provide well-tailored supports (e.g. reduction in other time-consuming responsibilities, remedial training, or additional information or resources) to relax those constraints as necessary to make the more demanding responsibilities feasible.

The paper proceeds as follows. Section I describes the context, intervention, and methodology. Section II presents the main impact estimates, while Section III reports the results of our mixed methods study of the program's theory of change. Section IV describes program costs. Section V ends by discussing the paper's contributions to the literature on teacher training program design and impact and its implications for how "good governance" is conceptualized in the literature on efforts to improve public service quality.

I. Research Design and Methods

A. Context

Students in Nepal complete basic education (grades 1 through 8), before progressing to secondary education (grades 9 and 10). At the end of grade 10, they take the Secondary Education Examination (SEE), which determines whether they can continue to upper secondary (grades 11 and 12). While primary enrollment rates have been high for several decades (with a primary net enrollment rate of 89.1% by 2008), enrollment at the secondary level has risen sharply in recent years In particular, the secondary level (grades 9-10) net enrollment rate increased from 35% in 2008 to 66% in 2017 (Ministry of Education 2018).

We focus on government schools that include grades 1-10 or 1-12, which account for 97% of Nepal's 9th and 10th grade students who are in government schools. According to our (nearly) nationally representative baseline data, most of these combined primary-secondary schools are relatively large: 88%

have 200 or more students, and 38% have 500 or more students. Many government schools are remote; the distance to the nearest motorable road is one hour or less for only 43% of schools, but more than three hours for 32%, and more than five hours for 17% of schools. Just 78% of schools have electricity, and only 32% have internet connections. Many families of secondary students are poor by high-income country standards, though they are not among the poorest of the poor; 95% of student families own mobile phones, 54% own televisions, and 40% own bicycles, but less than 13% own refrigerators and less than 10% own computers.

For secondary level teachers in Nepal's government schools, the required pre-service training takes three or four years and is conducted by universities (Gautam 2016; Pillay et al. 2017). Until 2018, in-service trainings were designed and overseen by the National Center for Educational Development (NCED) and held at 29 regional Education Training Centers. Prior to the trainings evaluated here, the most recent wave of government trainings took place under the School Sector Reform Program (SSRP) from 2009 to 2015. In principle, the SSRP required teachers to attend three 10-day training sessions over the course of five years. The training curricula were to be developed to address context-specific pedagogical needs as expressed by teachers and school officials. Using baseline data, however, we estimate that, in 2017, only 30% of secondary math teachers had received at least some SSRP math training, and only 21% of secondary science teachers had received at least some SSRP science training. Background documents for the SSDP (2016-2023) and preliminary qualitative research for this evaluation suggest that the SSRP trainings tended to be "too theoretical" and led to little if any real change in classroom teaching practices. SSRP evaluation reports also recorded the belief that the "demand based" approach to training content development caused training quality to suffer, both because teachers had difficulty articulating their needs and because the training institutions had difficulty designing quality content and hiring trainers with adequate skills (Poyck, et al. 2016). Baseline classroom observations using the Stallings tool (World Bank, 2015) indicate that while teachers (who know they are being observed) spend 76% of class time on instructional activities, the great majority (83%) of that time is spent using traditional teaching methods such as lecturing from the blackboard, rather than more interactive activities.[6]

B. SSDP Teacher Training Intervention

We study trainings that the Government of Nepal rolled out under the School Sector Development Program (SSDP) for teachers of 9th and 10th grade math and science in government schools. Given the perceived failure of the SSRP's demand-based approach to training curriculum development, the SSDP training curricula were developed in a more centralized way and covered challenging math or science concepts in the 9th and 10th grade student curricula, together with specific demonstration-based methods (often using teaching aids made from local materials) to teach specific concepts. Participants attended 10-day sessions

[6] We thank Rashmi Menon and Anuja Venkatachalam for help in implementing the Stallings observations.

at Education Training Centers (ETCs), and then were expected to complete five days of self-study project work that included: a) 10 lesson plans; b) an action research project related to a classroom or school problem; and c) two of several specified activities.[7] Teachers were to submit a report on the project work, approved by their head teachers, within 52 days of completing the ETC-based training. Most ETC sessions took place on school days. Teachers received per diems for their stays at the ETCs, but were not offered other monetary incentives to attend. They were also to receive grades for the training based on attendance, participation, test performance at the end of the ETC session, and project work. In principle, adequate scores were required for teachers to obtain credit for the training in their general performance review records.

The intervention we study invited all secondary math and science teachers in selected government schools to enroll in relevant SSDP trainings. The study intervention differed in small ways from the trainings that were to be rolled out nationwide. First, rather than wait for teachers and schools to request the trainings, the ETCs sent invitation letters to treatment schools (sometimes with follow-ups by phone), inviting them to send all secondary math and science teachers to attend the SSDP trainings. They did not invite teachers in control schools or in other schools in the same small geographic areas as the control schools. Second, while the broader roll-out prioritized teachers with permanent positions who were not trained under the SSRP, ETCs were asked to invite all teachers of 9th and 10th grade math and science in study schools, regardless of their employment status or previous training.[8] Third, the SSDP trainings were intended to include two modules, each with 10 days of ETC training and five days of project work, but in practice only the first module was rolled out. The national roll-out stalled shortly after our treatment took place, because of a dramatic "federalizing" reform that accelerated in September of 2018, when the ETCs were shifted from federal-level administration to administration by new provincial governments.

C. Sampling and Randomization

Our sample design was shaped by four objectives, which were developed through conversation with our government partners; they were to: (a) estimate student-level test score impacts with adequate precision; (b) generate descriptive statistics that are approximately representative of all of Nepal's combined primary-secondary schools (when using population weights); (c) direct study trainings primarily to teachers who

[7] For math training, possible activities were: (1) collect three-dimensional solids to use when teaching surface areas; (2) prepare a water tank model to study volume; (3) visit two local banks to obtain interest rate data for use in teaching compound interest; or (4) use a clinometer to calculate height and distance for objects near the school. The science training options were: (1) prepare a circuit and a related experiment for teaching electrical resistance; (2) create models of a human heart and a stethoscope and develop four related teaching exercises; (3) prepare a hydrocarbon model and methods to use it in teaching; or (4) prepare a planetarium model (on an umbrella) to teach about constellations.

[8] While this implies a difference between the study intervention and the SSDP intervention in non-study districts, we believe that the study remains relevant for Nepal's policy discussions. Baseline data reveal that 73% of grade 9 and 10 math and science teachers are in non-permanent positions, suggesting that policymakers will face important choices about whether, and how much, to invest in training non-permanent teachers. NCED records also indicate that there is precedent for government training of non-permanent teachers.

had not been trained under the previous SSRP training; and (d) derive impact estimates free of biases due to possible spillovers from teachers in treatment schools onto teachers in control schools. Conservative power calculations suggested the need to include about 100 treatment and 100 control schools.[9]

To obtain a sample representative of most of Nepal, we randomly selected 16 of Nepal's then 75 districts (after eliminating from consideration seven remote districts, two districts with administrative boundaries that crossed province borders and one district where data collection would be difficult for political reasons), and then sampled schools within those districts. Figure 1 shows the 16 selected districts, which are spread across all seven of Nepal's provinces. Within districts we sorted schools into two strata, "priority" and "non-priority," over-sampling the former. Priority schools were defined, according to rules that reflected the idiosyncrasies of the available records, as those for which the NCED's hard copy records showed no teachers with permanent or unknown contract type having completed training offered under the SSRP. Within each district, we selected two-thirds (8) of our sample schools from the priority stratum, and one-third (4) from the non-priority stratum. There were two exceptions to this sample selection scheme: (a) at the request of our government partners, we doubled the number of schools in the largest of the 16 districts; and (b) we selected only three non-priority schools in one smaller district since it had only three such schools. In total, our sample includes 203 ($12\times15 + 24 - 1$) schools. To limit spillovers from treated onto untreated schools, we first grouped schools (within districts) by the Village Development Committee (VDC) areas to which they belonged (before federalization). The average VDC had 1.7 eligible schools. We then sampled VDCs, and randomly selected only one school per VDC. Further details on sample design are given in Schaffner, Glewwe and Sharma (2018).[10]

Within each priority (non-priority) stratum, four of the eight (two of the four) schools were randomly allocated to the SSDP teacher training. Within cells defined by district and treatment status, we randomly allocated schools to two student assent processes and to two orders of assessment administration ("math first" and "science first").[11] Random selection was done without replacement, using a random

[9] We aimed for a sample large enough to give an 80% chance of detecting (at the 95%, two-tailed significance level) an intervention impact on average student test scores of at least 7 percentage points (about 0.3 standard deviations of the distribution of students' scores). Using an earlier academic achievement test, we estimated the standard deviation of the test score variable to be about 20, with a very high intra-cluster correlation coefficient of around 0.65. We expected to obtain greater precision by using baseline test scores as controls in endline impact regressions.

[10] Our design also randomly allocated half of the treatment schools to receive a small add-on "video assignment" intervention that was intended to raise teachers' motivation to complete the self-study project work. The effect of the add-on was far from statistically significant (see Schaffner, Glewwe and Sharma 2021) so we do not examine it here.

[11] At baseline, enumerators administered the informed assent process for students before distributing the assessments. After observing low baseline test scores, and concerned that the assent process may have made the low stakes nature of the assessments more salient to students, we administered the endline assent process to most students after they had taken the assessments but before they submitted their assessment papers (giving them an opportunity to choose not to submit). Following Institutional Review Board directives, whether conducted before or after the assessments, the assent process informed students that their scores would not count toward their grades in school and would not be revealed to anyone at their school. We retained the baseline assent process in one-third of the schools, hoping to

number generator in STATA 13. Teachers in the schools randomly allocated to treatment were to receive invitations to trainings in late 2017, while teachers in control schools (and neighboring schools in the same VDC) were to receive training only after the end of the study in May of 2019. Within schools, we invited all teachers of secondary math or science and attempted to gather data from all secondary students.

To produce estimates of the mean or variance of a population characteristic, or the average of a heterogeneous effect, for the population of schools in the 16 study districts, population weights are needed to adjust for differences in the number of schools per district and for district-specific population shares of priority and non-priority schools. We calculated these weights using Monte Carlo methods; when we use these weights, we interpret our sample as "nearly" nationally representative.[12]

D. Student Learning Outcomes

We first estimate the impact of the SSDP teacher trainings on student learning outcomes, as measured by scores on endline academic assessments. (We also estimate impacts on intermediate outcomes, which we describe in section III.D.) We administered two one-hour assessments, one in math and one in science to almost all 9th and 10th grade students present on the day of assessment. These endline assessments were drafted by U.S.-based consultants with expertise in psychometrics and familiarity with international assessment item banks, based on English translations of Government of Nepal student textbooks and SSDP teacher training curriculum documents. They were asked to give special attention to curriculum content emphasized in the SSDP trainings, to include items at lower grade levels (to assess how many students enter grades 9 and 10 with skills below grade level), and to include some items from the baseline assessments (to allow linking to those assessments).[13] They drew primarily from international assessment item banks, allowing incorporation of items that have been refined through intensive pre-testing. We used two assessment versions ("A" and "B") for each subject and grade, using common items to link their scores, to reduce the risk of copying among students sitting in rows in crowded classrooms and to increase the subject content covered by the assessments. The drafts were reviewed, amended, and translated into Nepali by local

evaluate whether the change in the assent process improved test performance. Procedures for asking teachers or head teachers to encourage students to do their best were strengthened at endline in all schools. We detected no impact on test scores of assent process timing, nor of the order of the tests, so we do not report these results in this paper. We nonetheless include test administration controls in our regression, to comply with our pre-analysis plan.

[12] Monte Carlo methods were used to account for the complex structure of sampling without replacement. The details of how these weights were constructed are reported in Appendix D of Schaffner, Glewwe and Sharma (2021). As it turns out, most weighted and unweighted estimates are very similar. The sample is not fully nationally representative because, as explained above, the sample frame excluded 10 districts (where only 6.1% of Nepal's government combined primary-secondary schools are located) and our weights do not adjust for small departures from using sampling probabilities proportional to size when selecting the 16 districts (see Schaffner, Glewwe and Sharma 2021).

[13] The baseline assessment items were developed by local experts under a tight time schedule. Subsequent analysis indicated that several items did not perform well, so we decided to form a joint team that combined these local experts with international experts to develop the endline assessments, drawing heavily on international assessment items.

assessment consultants, to ensure the relevance of the tests to the Nepali curriculum and testing style. After pre-testing, the six questions with the lowest correct response rates were dropped from each assessment, which produced the final assessments with 35 items each. All items are multiple choice.[14]

For each student we estimated, separately by grade, four achievement indices using Item Response Theory (IRT) methods to measure: overall math achievement, overall science achievement, achievement on a subset of math items most closely tied to the SSDP training content, and achievement on a similar subset of science items. For each grade and subject, we linked the two or three assessment versions by estimating a 2-parameter logistic IRT model (estimates shown in Online Appendix Tables A4–A7 of Schaffner, Glewwe and Sharma 2021).[15] We used item maps from the assessment consultants to identify the items closest to the content of the SSDP training.

E. Estimation methods

Our estimation methods follow our pre-analysis plan, which was registered with the American Economic Association RCT Trial Registry on June 24, 2019.

Student-level ITT impact regressions. For all eight student test scores described above, we estimate Intention to Treat (ITT) impacts of inviting all secondary math or science teachers in a school to the SSDP trainings. The main regression equation, which we estimate for all endline sample students, has the form:

$$Y_{is1} = \beta_0 + \beta_1 Treat_{s1} + A_s\beta_A + S_s\beta_S + \varepsilon_{is1} \quad (1)$$

where Y is a student learning measure, Treat is an indicator that a school was randomly selected for the trainings, A is a vector of indicators showing random allocation of schools to different assessment administration procedures (described above), S is a vector of district by priority/non-priority stratum fixed effects, i indexes student, and s indexes school. The subscript 1 refers to endline. Restricting attention to the "panel sample" of students present at both baseline and endline, we also estimate impact on endline scores while controlling for baseline scores, as in this equation:

$$Y_{is1} = \beta_0 + \beta_1 Y_{is0} + \beta_2 Treat_{s1} + A_s\beta_A + S_s\beta_S + \varepsilon_{is1} \quad (2)$$

where the notation is the same as above, and the subscript 0 refers to baseline. The main specifications are estimated by weighted least squares, using the weights described above. In all student-level regressions, we cluster standard errors at the school level, because the treatment was assigned at that level.

[14] Printer errors resulted in a small percentage of assessment copies with one incorrect page for the "A" versions of the 9th and 10th grade math assessments. The faulty assessments were distributed in a few schools before the problem was detected. We treat these as a third version ("C") of the relevant math assessments.

[15] Following our pre-analysis plan, we first tried 3-parameter logistic IRT models, but the estimates did not converge.

We estimated local average treatment effect (LATE) impacts, which are impacts on student test scores of teachers' actual participation in (rather than invitation to) the SSDP trainings, instrumenting teachers' participation in the trainings by their school's treatment assignment. As expected, the estimated impacts are larger in absolute value than the ITT estimates, but their statistical significance usually matches that of the analogous ITT estimates. See Schaffner, Glewwe and Sharma (2021) for our LATE estimates.

Estimating heterogenous impacts. Following our pre-analysis plan, we examined impact heterogeneity by adding (one at a time, in our ITT regressions) interaction terms between the treatment indicator and indicators of teacher characteristics (extent of the SSRP training, whether the teacher has a permanent contract, and whether the teacher has over five years of experience), a school characteristic (a school management quality index, described in Appendix A of Schaffner, Glewwe and Sharma 2021), and student characteristics (gender, whether at least one parent has a secondary education, a household wealth index, and student's tercile in the baseline assessment score distribution). We also include un-interacted heterogeneity variables. We found very few statistically significant interactions and concluded that we do not have power to detect heterogeneity along these dimensions. The estimates are reported in Schaffner, Glewwe and Sharma (2021). We also estimated quantile regressions to examine possible heterogeneity by quantile of the unconditional endline assessment score distribution, but again we concluded that our estimates were too imprecise to draw clear conclusions.

Multiple hypothesis testing. We calculate False Discovery Rate adjusted p-values following Benjamini and Yekutieli (2001). We do this for tests of no differences in outcomes between treatment and control schools for all ITT estimates of the program's impact on students' endline test scores, separately for estimates using all endline students and using students with panel data, and separately for scores based on all test items and based only on items closely related to the SSDP training.

F. Broader mixed methods data collection

To evaluate the assumptions underlying the SSDP training program's theory of change, we embedded the randomized evaluation of the program's primary impacts within a larger mixed methods study that included analysis of monitoring data, a part-qualitative and part-quantitative telephone survey of trainers and training participants, a small-N qualitative study, and quantitative analysis of intermediate outcomes.

Due to concerns about weak government practices regarding collection of data to monitor the trainings, we arranged for research assistants to collect monitoring data through frequent phone calls to the ETCs. They gathered data on training session dates, reasons for training session delays, compliance with protocols on who should and should not be invited to the trainings, numbers of trainers, teacher invitation activities, and teacher attendance (see Shrestha 2019 for details). These calls also elicited qualitative

information on the difficulties ETCs experienced in rolling out the trainings. We do not believe that our collecting the monitoring data significantly increased the sense of accountability among ETC personnel.[16]

To obtain more detail on the quality and de facto content of the SSDP trainings, and the institutional context shaping decisions by trainers and trainees, we sought to interview by phone all teachers from the baseline sample who had been in treatment schools at baseline and had completed the SSDP trainings. Of the 221 secondary math and science teachers in treatment schools, we have baseline data for 192. Of these, 120 completed SSDP training, and we were able to interview 98 (representing all 16 study districts) by telephone. Seeking to interview one math trainer and one science trainer in each of the 14 ETCs serving our 16 districts, we succeeded in interviewing 12 math trainers and 11 science trainers. The interview protocols included both closed-ended and (short answer) open-ended questions. They focused on the content, methods, and quality of the trainings, and on whether teachers were held accountable by any stakeholders in their schools for reporting on the content of the trainings or for implementing new teaching practices after training. For more details, see Schaffner, Glewwe and Sharma (2019).

To confirm and deepen our understanding of the telephone interview results, we commissioned a small-N qualitative study (Acharya and Upreti 2019), which was implemented in three of the 16 districts in the quantitative study, one each in Nepal's Eastern Terai, Western Mid Hill and Far Western regions. In each district, two math and two science teachers who had completed the SSDP trainings, plus another teacher (a local teachers' federation representative), were interviewed. In each district, one trainer involved with the SSDP math training and one involved with the SSDP science training were also interviewed.

Finally, we gathered quantitative data at baseline and endline that would shed light on additional elements of the theory of change. For example, we collected data on school management structures and practices at baseline, and at endline we measured intermediate outcomes on teacher subject knowledge, attendance and teaching practices for all teachers, and asked head teachers and teachers in treatment schools about teachers' completion of the self-study project work associated with the SSDP trainings.

II. Student Learning Impact Results

A. Implementation Fidelity, Attrition and Balance

Monitoring data gathered through phone calls with ETC personnel confirm that the ETCs succeeded in inviting all treatment schools to send their 9th and 10th grade math and science teachers for the SSDP trainings, and in preventing teachers from control schools (and other schools in the control schools' VDCs) from attending the trainings during the study period. All ETCs rolled out both the math and science

[16] The ETC personnel that our team spoke with did not seem to consider themselves accountable to our team. In fact, repeated phone calls were needed to obtain even basic data, and one ETC provided no information.

trainings, though more slowly than initially planned. In the treatment sample the mean exposure to the math training treatment was 11 months (from the end of the SSDP training session until endline data collection), and to the science training was 9.4 months.[17] We discuss these delays, as well as evidence on the de facto content and quality of the training sessions, below.

Overall student attrition rates between baseline and endline, shown in Table 1, while high (35-36% in grade 8 and 41-44% in grade 9), are similar in treatment and control schools, differing by only 0.5 percentage points for grade 8 and 2.9 percentage points for grade 9. The most common reason for attrition was being absent on the day of the endline test, followed by no longer being in school. For further details on attrition, see Schaffner, Glewwe and Sharma (2021). For grade 8, attriters' average math test scores at baseline are (marginally) significantly lower in the treatment group than in the control group. Thus, as long as any larger impact of the SSDP program on attriters relative to its impact on non-attriters is smaller than this baseline gap, including attritters at endline would tend to reduce test scores in treatment relative to control schools, reducing estimates of SSDP training impact on test scores. This would strengthen the results below, since most SSDP impact point estimates are negative. For grade 9, the difference in the attrition rates, while marginally statistically significant (p-value = 0.065), is small (2.9 percentage points), and there is little difference between treatment and control in baseline test scores among attriters.

Regarding balance across the treatment and control schools, Tables 2 and 3 show little evidence of imbalance at baseline. Overall, of the 29 school, teacher, and student variables in these tables, two are significant at the 5% level and another two at the 10% level, which is close to what one would expect from random chance even if all true differences were equal to zero. Note also that the absolute sizes of the differences are not large, and they lose significance after adjusting for multiple hypothesis testing. Below we check the robustness of our main findings to including controls for the significant variables. Following our pre-analysis plan, we also check robustness by including controls for key school-, teacher-, and student-level variables of possible interest for studying impact heterogeneity.

B. Student Learning Impact Estimates

Table 4 presents ITT impact estimates for the eight endline student test score variables, disaggregated by grade (9 or 10), subject (math or science), and two scores for each assessment: a total score using all 35 questions and an "SSDP focus" score using questions most closely tied to the SSDP training. All scores are normalized by subtracting the mean and dividing by the standard deviation of the control group distribution.

Our main ITT regressions, using the full endline sample (Table 4, third column), yield no evidence that the SSDP teacher trainings increased student test scores. Six of the eight estimates are negative, of

[17] The exposure to the math training ranged from 9 to 14 months, while the exposure to the science training ranged from 7 to 15 months. Though less than ideal, this was enough time for most teachers to implement improved teaching methods across most of the 9th or 10th grade curriculum content, and so enough time to affect student learning.

which three are statistically significant at the 10% level. The 95% confidence intervals rule out effects above 0.10 standard deviations (of the distribution of test scores) in six out of eight cases and rule out effects above 0.18 standard deviations in all cases. ITT regressions on the panel sample, with baseline test scores as controls (Table 4, fifth column), largely confirm these results. Adding the baseline scores reduces slightly the standard errors of the estimates, ruling out positive effects at slightly lower thresholds.

Following our pre-analysis plan, we checked our results' robustness in five ways. Results are very similar if we: (1) use unweighted instead of weighted regressions; (2) omit test-taking control variables; (3) add school, teacher, and student controls; (4) add controls for variables not balanced at baseline, or (5) use normalized raw percentage scores instead of normalized IRT-based indices as the dependent variables. For details, see Supplementary Table 2 in Appendix F of Schaffner, Glewwe and Sharma (2021).

The robust conclusion is that the SSDP teacher training was unsuccessful in producing a sizeable increase in student learning. The following section investigates in detail possible reasons for this failure.

III. Weaknesses in the Program's Theory of Change

Before developing our mixed methods research design, we worked with our government partners to articulate the intervention's theory of change (which is summarized in Appendix Figure A1). Our approach to articulating this theory of change highlights the roles of trainers, teachers, and students, who are the key actors along the logical chain connecting the policy to student learning. To work out the theory's details, we first identified how these actors must respond to the new responsibilities and opportunities created by the SSDP policy for that policy to deliver the intended improvements in student learning. In each of these response areas, we identified what the actors would require in (a) motivation, (b) resources (e.g. finances, supplies, time) and (c) capacity (i.e. knowledge and skills) for them to respond in desirable ways, thereby making the program successful in raising student learning. We then designed our mixed methods evaluation to look for evidence of deficits in these areas, drawing on the four complementary types of data described above: monitoring data gathered directly from ETCs, telephone surveys with trainers and teachers who attended the SSDP trainings, small-N qualitative data from in-depth in-person interviews, and data from the baseline and endline quantitative surveys. In what follows we document five violations of the assumptions underlying the theory of change, which appear to offer important insights into the disappointing impacts of the SSDP training program.

A. Weak governance of ETC training sessions

For the SSDP trainings to improve student learning, program designers and administrators must provide good governance for the roll-out of ETC training sessions. More specifically, they must ensure that trainers have: adequate knowledge and skills for developing and delivering high quality training content; adequate preparation time, materials, and facilities; and adequate motivation to invest time and effort into helping

teachers improve their teaching practices. The motivation may be internal, arising out of trainers' desire to do a good job, or external, arising out of the knowledge that they will be held accountable by others. We find that, while trainers were sufficiently motivated to roll out training sessions of the required length (10 days) in all ETCs, the institutions that might have supported them failed to provide them adequate guidance, time or training for training session preparation, and the broader institutional environment dampened their hopes of effecting real change in teachers' practices.

All ETCs in our study districts rolled out trainings that lasted the full 10 days, although their rollout was somewhat later than intended. Initially planned for October and November of 2017 (shortly after baseline), no trainings began before December 2017 and most took place in April or May of 2018. These late starts were often due to delays in the earmarking and disbursement of government funds by the Ministry of Finance and the Ministry of Education, Science and Technology. Soon after the roll-out, when government funding limits were tighter than anticipated, we asked the ETCs to prioritize the math trainings, so that we could study one common training across all study districts, even if funding did not arise for further trainings. Twelve of the 14 ETCs complied by implementing the math training first. Additional funding eventually allowed all ETCs to roll out both trainings. For more details on the roll-out, see Shrestha (2019).

Trainers appear to have carried out their assignments with at least moderately good intentions. They spent time creating materials for full 10-day training sessions, and in telephone interviews they demonstrated significant engagement with their assignment, for example by identifying the ways they wanted teachers' practices to change. According to ETC personnel reports, adherence to the NCED curriculum guidelines for the training sessions (distributed in 11-page documents) varied across ETCs (Shrestha 2019). Most of the 13 ETCs for which we have data used variants of the SSDP curricula for both math and science, but two (three) used a pre-existing curriculum for the math (science) trainings, reporting that they did not receive the guidelines soon enough or were unaware that new guidelines existed.

Post-training telephone interviews with participants and trainers suggest that training content was broadly consistent with the curriculum described in official documents, emphasizing practical methods (often involving teaching aids made from local materials) for teaching specific secondary level math and science concepts. To identify the math or science topics that teachers found most memorable, we asked teachers to list up to four "math or science concepts or skills in the secondary curriculum that were explained, discussed or practiced during the training," focusing on the topics to which the most time was devoted. The most frequently mentioned math domains were mensuration (with mentions of activities such as making cylinders from pieces of paper to help students learn to calculate surface areas) and trigonometry (often referring to use of a clinometer). For science topics, the domains most often mentioned were biology (with mentions of bringing plants to class when teaching about roots, stems, and leaves) and chemistry (with more varied examples). Answering other open-ended questions, teachers mentioned making and using

litmus paper and using seating arrangements to teach about sets or the periodic table of elements. In some, but not all, ETCs the trainings also emphasized the pedagogical practice of having students work in groups.

The data suggest, however, that the quality of the trainings fell far short of the ideal. In telephone interviews, 19 out of 23 trainers mentioned that they received no training to conduct the SSDP trainings, and many mentioned that the curriculum documents lacked adequate detail and arrived too late to prepare for the training sessions. In open-ended responses, training participants reported that trainers seemed inadequately prepared. In addition, some trainers' attitudes may have diminished their training efforts: 61% mentioned inadequate teacher motivation, poor teacher attitudes and/or poor monitoring of teachers when asked what obstacles might prevent teachers from implementing new teaching methods.

When asked to rate trainers' performance directly, many teachers responded positively, perhaps out of respect or politeness; over 60% rated their trainers as very good, and 9% rated them as excellent. Yet teachers' answers to more open-ended questions revealed significant discontent with the performance of the trainers. When asked what they most disliked about the trainings, one fifth (20%) of teachers expressed displeasure with trainer content knowledge or preparation. When asked what problems reduced the trainings' effectiveness, 14% of teachers mentioned that the trainers arrived late on training days, and 11% mentioned lack of skilled trainers or subject experts. In open-ended responses about how to improve the trainings, 31% of the teachers said that trainers should be content experts, and 8% said that the trainers should be trained better. Teachers stated that sometimes when they asked questions of clarification, trainers could not answer them, and sometimes trainers could not solve problems that teachers were asked to solve.

While the ETCs' facilities, equipment and supplies seemed adequate for standard lectures and small group discussion, in some cases they were inadequate for practice with lab experiments, classroom methods or information technology (IT) tools. When asked what prevented trainings from being more effective, 15% of teachers mentioned inadequate materials and 11% cited inadequate facilities, including lack of labs and IT equipment. Several trainers also referred to lack of materials or poor facilities (such as lack of electricity) as problems for the trainings. When asked how to improve trainings, 18% of teachers mentioned the need for more teaching materials and 35% mentioned that the training center needed a lab. Several trainers also suggested that trainings could be improved by having adequate facilities and materials. (Also, 9% of teachers cited a need for better accommodation.) Given the SSDP's emphasis on equipping teachers to use materials and demonstrations when teaching, the lack of materials and lab facilities is especially worrisome.

Despite these weaknesses in the roll-out of the SSDP training sessions, attendees came away with ideas about how to apply new methods in their classrooms. When asked "On the basis of what you heard or learned during this training, what changes do you most hope or expect to make in the way you teach secondary math or science?", nearly two thirds (64%) of teachers (of both subjects) mentioned using materials, demonstrations, or experimental methods. A small percentage (13%) also said they intended to

use group work more. Overall, teachers were generally positive about the trainings, though their responses suggest much room for improvement. Over 80% of interviewed teachers reported that they found at least half the content to be very valuable, but only 40% of science teachers and 17% of math teachers found at least three-quarters of the content to be very valuable.

B. Significant teacher non-participation

If trainings are to improve student learning, teachers must attend them. Unfortunately, participation in the SSDP trainings was disappointingly low. The teachers to whom students were most exposed during the treatment period were those present in the schools at endline. Only 60% of these secondary math teachers and 42% of the science teachers had completed the SSDP trainings. Moderately high teacher turnover rates meant that some teachers were untrained because they had joined the schools after the trainings took place, yet the most important cause of low participation is that many teachers present in the schools at the time of the training did not participate. Among math and science teachers in treatment schools at baseline, 20% and 29%, respectively, had left the schools before the endline. Correspondingly, 18% of math teachers and 25% of science teachers in the schools at endline were new since baseline. Even among teachers in the schools from baseline through endline, however, only 68% of math teachers and 52% of science teachers participated in the SSDP trainings.

Difficulties in finding qualified substitutes to cover teacher's classes reduced the participation rate. During monitoring calls, ETC personnel reported that some schools refused to send teachers to the trainings for this reason. Two observations from the baseline data suggest why this may be important. First, 78% of schools have only one secondary math teacher and 83% have just one secondary science teacher, suggesting difficulty of finding other teachers from the same school to cover secondary level math and science classes. Second, over one-fifth of head teachers reported that when teachers leave for training, the teachers' classes are cancelled or left unsupervised. A recent study commissioned by the Ministry of Education highlights a broader shortage of adequately prepared math and science teachers (Pant *et. al.*, 2018).

We suspect, however, that some teachers had permission to attend, but opted not to do so. Some may have thought that the training would be very similar to previous trainings. We cannot quantify this directly, but several training participants reported arriving with little expectation of learning anything new (based on past training experiences) and being pleasantly surprised to learn new skills. Others declined the invitation for personal reasons, such as the need to prepare for teacher examinations (an unusual opportunity that arose during the study period, through which temporary teachers could gain permanent status), travel, illness, or family obligations. A few participants cited inadequate per diems or accommodations, suggesting that the costs of participating may have dissuaded some other teachers.

C. Mismatches between training content and teachers' subject knowledge

For trainings to succeed in improving teaching and learning, participating teachers must master new knowledge or skills during the trainings. Unfortunately, some teachers who participated in the SSDP trainings appeared to lack math and science subject knowledge that they would have needed to fully benefit from the SSDP trainings, which focused on advanced math and science concepts. To assess teachers' subject knowledge without explicitly asking them to take assessments, we asked (at endline) for teachers of 9th and 10th grade math and science to fill out anonymous evaluations of selected student assessment items. The evaluation forms presented 12 items from the student assessments and asked teachers to: (1) rate each item's clarity; (2) provide the answers they thought the items' writers intended as the correct answers; (3) estimate what fractions of their students would answer each item correctly; and (4) rate how well-tailored each item was to the Nepalese curriculum. Teachers were surprisingly willing to complete the evaluations. We concluded that one item each in the math and science assessments were poorly designed. Teachers' responses to the remaining 11 questions in each subject are reported in Tables 5 and 6. One fifth of math teachers gave incorrect answers to question 9 (Table 5), a straight-forward algebra problem. Given that the SSDP training was supposed to deal with more advanced algebra techniques, such as factoring polynomials, teachers who are weak in basic algebra may find the training content difficult to follow. More concerning, nearly 40% of responding teachers gave incorrect answers to question 8, which requires understanding of how to calculate the perimeter of a rectangle. Again, having difficulty with this basic problem may prevent a teacher from understanding the more advanced SSDP training discussion of surface area calculations for 3-dimensional solids. One fifth (20%) of science teachers gave incorrect answers to question 4 (Table 6), which asks about the main function of red blood cells and is likely to be an important building block for the SSDP's focus on the circulatory system. Nearly half (48%) incorrectly answered question 8, concerning evaporation; this is likely foundational for SSDP training content on climate change.

To estimate the impact of SSDP trainings on teachers' subject knowledge, we used IRT methods similar to those used with student test scores to create teacher subject knowledge test scores. Table 7 reveals little or no impact of the SSDP trainings on teacher subject knowledge; see Appendix A (Tables A17-A19) in Schaffner, Glewwe and Sharma (2021) for details. Average standardized teacher test scores are *lower* in the treatment group than in the control group, though the differences are not statistically significant.

D. Weak motivation and support for post-training changes in teaching practices

Even when trainings succeed in increasing teachers' knowledge or skills, they may have little impact on teaching and learning if teachers lack adequate motivation for making the daily investments of time and energy required for teaching in a new way, or if they lack adequate guidance for this work. We find that the time investments required for adopting SSDP teaching practices were probably daunting to most teachers, that the typical school environment provides teachers with little accountability for adopting new practices,

and that lack of resources for teaching materials may also have inhibited adoption. Unsurprisingly, our quantitative study finds no evidence that the SSDP training program had any impact on teaching practices.

To transform their practices in the ways promoted by the SSDP trainings, teachers must write many lesson plans incorporating new demonstration-based methods and must prepare visual aids. Investing the time required is probably an especially daunting challenge for teachers in Nepal's government schools, who are accustomed to teaching without much preparation. Nearly half of secondary math and science teachers (41%) reported (in baseline data) preparing for 15 minutes or less for each class period they teach and only nine reported (in endline data) using written lesson plans to guide their classroom teaching.

Baseline data suggest that schools often lack leaders who hold teachers accountable for how they teach, much less for adopting new teaching practices after training. SSDP trainers had no responsibilities for interacting with teachers after the SSDP training sessions. Parents, most of whom do not have secondary education, are ill equipped to monitor the quality of secondary teaching. School Management Committees, while often active in improving facilities and equipment, tend to provide little or no oversight over what happens in classrooms, leaving the day-to-day management of schools to the head teachers (Schaffner, Glewwe and Sharma, 2018). While, in principle, School Resource Persons (SRPs) had responsibilities for supporting good pedagogical practices during the study period, in practice each SRP was responsible for 28 schools on average, and they were spread too thinly to provide adequate support (Ministry of Education 2016). Nearly half (49%) of teachers report having never been observed even a short time by an SRP or school supervisor in the past year, while another 48% reported only 1-4 such visits and 3% reported 5 or more visits (baseline data). The actors most likely to hold teachers accountable for change are head teachers. Most head teachers, however, are very busy teaching, with the median head teacher carrying a class load that is almost three fourths (72%) of what a standard teacher carries. This leaves them with little time to observe other teachers; in 65% of schools, the median teacher reports that the head teacher "never" observes them for a whole class period. Head teachers appear even less likely to hold teachers accountable for adopting new practices after training. In telephone interviews, most trained teachers (85.7%) reported that they were requested (by their head teacher or colleagues) to report on what had happened during the trainings, but only a few (11.2%) reported being asked (by anyone) to report on how they were teaching differently because of the training, and none reported being *required* to report on the training in any way.[18]

Reports by head teachers and trainers suggest that teachers indeed lack adequate motivation for adopting new practices. According to head teachers' reports at endline, only 34% of the teachers who

[18] We also know from a qualitative study of a small add-on intervention (in which teachers were visited by training personnel who were to video-record the teachers teaching full class sessions to demonstrate their adoption of new teaching practices) that at least some teachers welcomed the opportunity for monitoring and feedback. Other teachers volunteered that the SSDP trainings could be improved by adding monitoring and follow-up after trainings.

attended SSDP trainings were highly motivated to try new teaching materials or methods after the training. When asked their hopes of how teachers would change their teaching practices after training, most trainers said they hoped teachers would use "practical" or "experimental" methods, demonstrations, or low-cost teaching materials. But when asked how likely it was that teachers would change their teaching in these ways, only one of 23 trainers said "very likely", while just over half responded "somewhat likely" and 43% responded "somewhat unlikely." When asked about the main obstacles to adoption by teachers, 61% of the trainers mentioned inadequate motivation, poor attitudes and/or lack of monitoring. Some mentioned especially that teachers are unmotivated to introduce new teaching materials because they teach many classes and do not have enough time to prepare and implement new methods. More tellingly, teachers apparently lacked accountability even for the narrower, short-term task of completing the self-study project work component of the SSDP teacher training. Among teachers who attended the SSDP trainings, their head teachers reported that 37% had not completed the lesson plan development assignment. Moreover, the SSDP curriculum requires teachers to complete one of three or four specific additional projects, but when we asked trained teachers at endline to name which of these projects they had completed, very few could even name any of the options. Of the math teachers who attended SSDP trainings, only one could describe a project similar to one of the official project options, while 26% said they did not do these projects and another 58% said they did not know or could not recall the projects they did. The findings for science teachers are similar. Since the unusual projects should have been memorable, we conclude that it is likely that few teachers gave serious consideration to doing the self-study project work. One teacher interview gives further reason to believe that there was little accountability for doing the self-study project work well: "I developed lesson plans and did project work. The Resource Person stamped [my paper] but nobody checked it." Thus, we suspect that few teachers completed this potentially important part of the training. Institutions that fail to hold teachers accountable even for completing these small projects are unlikely to hold them accountable for the larger investment needed to adopt new practices on an on-going basis.

Lack of teaching materials may also have limited the adoption of new teaching practices. When asked "What obstacles have you run into, if any, when trying to apply what you learned during the training in your own classrooms?", 59% of teachers cited lack of teaching materials. The great majority (78%) pointed more broadly to lack of either materials or facilities, including lack of (14%) or poorly equipped (6%) labs, lack of information and communication technology infrastructure (18%) and other infrastructure problems (19%). Other obstacles cited by teachers included lack of teacher time (17%), too many students or difficulty managing students (15%), unmotivated students (7%), and low student attendance (5%). Interestingly, lack of adequate understanding or guidance was not cited as a barrier to adoption of new methods, though five teachers volunteered interest in having follow-up after training. An anecdote reported

during qualitative research suggests that teachers may also have found it difficult to implement the new demonstration-based teaching methods in class periods that are only 40 to 45 minutes long in most schools.

Unsurprisingly, we find little impact of the SSDP trainings on teacher attendance or teaching practices. We measured teacher attendance in three complementary ways: direct enumerator observation of teacher presence on the first day of a school visit, and student and head teacher reports regarding teachers' frequency of attendance. We measured teaching practices using head teacher and student responses at endline.[19] Given the focus of the SSDP training curriculum, we developed teaching practice questions around the use of written lesson plans, use of demonstration-based teaching methods, and efforts to incorporate local materials and information into classroom and homework activities. Given teachers' exposure in some ETCs to using group work as a teaching method employed by the trainers, we also attempted to measure teachers' use of group work. We also included practices not emphasized in the training curriculum, such as practices around the use and grading of homework. Many teaching practice outcomes are ordinal, with two or more categories for Likert-scale opinions or activity frequencies. We use probit (ordered probit) models to estimate ITT impacts on binary (multinomial) outcomes, while controlling for test-taking conditions, employing population weights, and (where relevant) clustering standard errors at the school level. For polychotomous outcomes, we first collapse categories (collapsing small categories into the adjacent category closer to the "middle" score) if categories have less than 5% of all observations.

Tables 7, 8 and 9 summarize our estimates of SSDP training program impacts on teacher attendance and teaching practices. The differences in teacher attendance between the treatment and control groups are small and statistically insignificant, whether measured by enumerator observation (Table 7) or by head teacher or student reports (reported in Schaffner, Glewwe and Sharma, 2021). We generally find small and statistically insignificant differences in teachers' teaching practices between the treatment and control groups, from both student (Table 8) and head teacher (Table 9) reports,[20] but there are three exceptions. Two of the three exceptions – head teacher reports on teachers' frequency of using teaching materials or visual aids and teachers' frequency of having students work in small groups – involve dimensions of teaching practice that are most likely to be affected by the SSDP teacher trainings. The third exception, whether a teacher required students to work on longer-term projects, is not obviously related to the SSDP training. Multiple hypothesis testing adjustments eliminate all three statistically significant differences. Comparisons of descriptive statistics for all these outcomes suggest, though, that even if some differences were statistically significant, they are not very large, with the probabilities of teachers employing practices at all, or frequently, differing between treatment and control groups by only a few percentage points.

[19] Estimated impacts based on teachers' self-reports were insignificant; see Schaffner, Glewwe and Sharma (2021).
[20] Teacher self-reports of teaching practices are also statistically insignificant; see Schaffner, Glewwe and Sharma (2021).

E. Mismatch between training content and many students' preparation for secondary school

For teacher training to increase student learning, it must improve the teaching of subject content that is appropriate to students' current level of understanding. Our data, unfortunately, raise major concerns about students' preparedness to learn the advanced 9th and 10th grade math and science concepts emphasized in the SSDP curriculum. At baseline, 93% of head teachers reported that a major challenge to teaching and learning in grades 9 and 10 was that students often enter these grades with below-grade-level knowledge. At endline, therefore, we included assessment items to test students' grasp of material that they should have learned in earlier grades and is foundational for learning the advanced concepts in the SSDP curriculum.

Tables 10 and 11 illustrate that many students have trouble with pre-requisite concepts. For example, consider Question 3 in Table 10, which local experts associate with the grade 8 mathematics curriculum in Nepal. At endline, 56% of grade 9 students and 46% of grade 10 students answered this incorrectly, suggesting that they had not fully learned or retained what they should have learned in grade 8. More worrying, over one fifth (21%) of grade 9 students failed to correctly answer Question 6 ("What is 6 divided by 3?"), which tests basic division facts at approximately the grade 3 level. The results for Question 8, which 33% of grade 9 students answered incorrectly, are also troubling. While this question is more difficult, requiring students to convert a narrative into mathematical symbols, it requires only basic math knowledge. Table 11 raises even more concerns about students' preparation for grade 9 and 10 science; for three of the eight questions, only about half or fewer of 9th and 10th graders answered correctly.

To address concerns that students answered questions incorrectly because they did not take the assessments seriously (rather than because they did not know the material), we use enumerator team reports to identify schools where: (a) the head teacher or teacher encouraged students many times to put effort into the assessments; (b) enumerators rated the students as taking the assessments at least moderately seriously; and (c) enumerators reported little or no distracting behavior during the assessments. When we limit the analysis to these schools, we obtain similar results; for details, see Schaffner, Glewwe and Sharma (2021).

IV. Program Costs

Our figures for SSDP program costs are rough approximations at best, because the trainings were rolled out by government institutions that use shared administrative capacity to roll out many programs, and because we had access to few detailed accounts. We began by listing all the activities required to develop the training curriculum and roll out the trainings, dividing them into two categories: (1) costs incurred only once for the entire country; and (2) costs incurred each time an ETC rolls out a training session for approximately 20 teachers. With a partner organization, Nepal's Center for Policy Research and Consultancy, we developed rough estimates of the time required to complete each task in the first category and the pay grade at which that task would be done, and used government pay scales to estimate cost. We estimated costs in the second

category from actual training session costs in one ETC. Note that international organizations working with education policymakers in Nepal, such as the Asian Development Bank, UNICEF, and the World Bank, also assisted in the design of the SSDP trainings, and we have not included the costs of their contributions.

The costs worked out to approximately $130 per teacher. The cost per student of training their teachers in one subject is about $2.60 ($3.00) per grade 9 (grade 10) student. By far the largest cost is the per diems and lodging for participating teachers. For details, see Schaffner, Glewwe and Sharma (2021).

To put these costs into perspective, we turn to Damon et al. (2019), who present cost information for some of the most effective interventions in their review, which increase average test scores from 0.2 to 0.4 standard deviations. The costs per student per year, excluding administration costs, range from $15-24 for a girls' scholarship program, private school vouchers, and a relatively costly computer-based program, to $3-9 for student incentive programs, to $2-3 for teacher incentive programs. Overall, while the SSDP intervention has a relatively low per-student cost, that cost, when used for other education interventions in other contexts, has generated large increases in learning. This highlights the need to improve the SSDP program's effectiveness or replace it with interventions that, for the same cost, can raise student learning.

V. Discussion and Conclusion

Recognizing the critical role that teachers play in determining how much students learn (Das et al. 2007, Clotfelter et al. 2010, Chetty et al. 2014), governments in low- and middle-income countries seek to improve student learning by providing teachers with in-service training. Learning how to design teacher training programs that are cost-effective for improving student learning is especially important, given the on-going "learning crisis" in those countries (World Bank, 2018). A small but growing literature seeks to illuminate good training program design with rigorous evaluations of teacher training programs in low- and middle-income countries; we identified 23 such studies.[21] Our research contributes to this literature in three ways.

First, our study supports a growing consensus regarding the need to re-design government teacher training programs of a common form, which bring groups of teachers to centralized training venues; instructs them in pedagogical theories or strategies but provides them with no lesson plans, teaching materials or other structured curriculum support; and includes no subsequent monitoring or coaching.[22] Studies of such trainings for secondary math teachers in China (Loyalka, e al. 2019), migrant school English

[21] See the Appendix Section B for a description of our approach to identifying the literature. Appendix Table B1 provides basic information on the 23 studies that we identified.

[22] Popova et al. (2019) show that about three fourths of teacher training programs in developing countries take place in a centralized location, only about half provide teachers with lesson plans or videos, only about one third provide scripted lessons, and only about one third provided follow-up visits for in-class pedagogical support or for monitoring. This stands in contrast to recent recommendation by the Global Education Evidence Advisory Panel (World Bank and FDCO, 2020), which highlights "structured lesson plans with linked materials and ongoing teacher monitoring and training" as one of only six "good buys" for improving education in developing countries.

teachers in China (Zhang, et al. 2013), and secondary entrepreneurship teachers in Rwanda (Blimpo and Pugatch 2021) find no evidence of impact, and in the first and third cases rule out anything more than small positive impacts.[23] Our study increases the list of null results to four, for this form of government training that is common but seldom evaluated rigorously.

Second, our qualitative evidence offers additional support for, and new hypotheses regarding, two emerging themes in the literature. The two themes highlight the importance for training program effectiveness of two corresponding design features: providing teachers with lesson plans and related teaching materials (or other structured curriculum support) that help teachers integrate new teaching approaches into all relevant course material throughout the school year; and the importance of following up initial group trainings with individual monitoring, coaching or other longer-term contact between teachers and training program personnel. Policy documents drawing on the growing evidence base promote these design features as best practices (Beteille and Evans 2019, World Bank and FDCO 2020). Perhaps because programs with such features are seen as most promising, most teacher training interventions for which we identified rigorous evaluations include variants of one or both features.[24] Evaluations have found positive and significant impacts on student learning in at least one subject for almost all these interventions. Acknowledging that the relative success of these programs may be explained by design features other than those highlighted, several studies have attempted to estimate the improvements in program impact arising out of the addition of these features. Albornez et al. (2020) find that, among seventh-grade science teachers in Argentina, the addition of structured curriculum support to standard government training greatly increases the impact, while the further addition of longer-term coaching brings a modest additional increase in impact. Among the primary school reading teachers studied by Cilliers et al. (2020a, 2020b), training with structured curriculum support substantially improved student learning, and the addition of coaching nearly doubled the impact early on, although both these results appear to fade over time. Among the first and second grade teachers studied by Piper et al. (2018), adding structured lesson plans to an intervention that included teacher training, coaching and revised textbooks doubled the impact of that intervention on student learning in math, Kiswahili and English. Policy documents, such as World Bank and FDCO (2020), assume (without presenting evidence) that both structured curriculum support and longer-term follow-up

[23] While the intervention evaluated in Lehrer et al. (2019) includes no long-term follow-up or structured curriculum support, we do not include it in the list of government trainings of the common form described in the text, because (unlike the more common government training programs) it provides a technology package of interactive whiteboard, digital content, and a computer projector.

[24] As explained in Appendix B, we defined structured curriculum support broadly to include provision of lesson plans and other materials that integrate new teaching practices into class sessions throughout the school year. We defined longer-term follow-up to include periodic monitoring and coaching (or the distribution of multiple training sessions over one or more school years), through which teachers and their practices are observed after initial training sessions. We identified eight interventions that included both elements, four that include only structured curriculum support but not longer-term follow-up, and seven interventions that included only longer-term follow-up (see Table B1).

are effective because they address teachers' skill gaps: the guidance provided by structured curriculum support is portrayed as substituting for missing teacher skills, while longer-term support is portrayed as helping to reduce skill gaps. Yet much remains unknown about the mechanisms through which these design features raise training program impacts and how to design these features in the most cost-effective manner.

Our study suggests additional possible channels through which structured curriculum support and longer-term follow-up might improve training program impact on teaching practices and student learning. Our qualitative evidence suggests that the time and resource costs of creating daily lesson plans (and related teaching materials) that embody the new teaching methods presented during the SSDP trainings inhibited teachers from adopting those methods. In such a context, providing structured curriculum support might increase adoption of new practices by reducing these costs, even among teachers with adequate knowledge and skills to create the plans and materials themselves. Our evidence also shows that teachers' classroom practices are largely unmonitored by parents, school management committees, head teachers, or School Resource Persons, implying that teachers are largely unrewarded by others for teaching innovation. In such a context, periodic visits after centralized trainings by monitors, coaches or other training system personnel may let teachers know that their efforts to adopt new practices will be noticed and rewarded. Even where public education system institutions fail to define or enforce any ties from teaching quality to pay or promotion, it may be possible to design follow-up training program activities in ways that increase the social or psychological rewards for pedagogical innovation. Examples of the effectiveness of such rewards have been shown by Ashraf, Bandiera and Jack (2014), who found that public health agents in Zambia who were given "stars" for performance and a public ceremony for top performers sold twice as many condoms as agents who were offered small monetary bonuses for each packet sold, and by Gauri et al. (2021), who found that a similar type of social recognition of health facility workers improved record keeping in one Nigerian state, though not in another. In addition, when teachers are time constrained and unmonitored, the provision of structured curriculum support and longer-term follow-up are likely to be complementary, since social and psychological rewards, which may be small, are more likely to outweigh the costs to teachers of adopting new practices when those costs are kept low by the provision of structured curriculum support.

Third, our study demonstrates how well-planned mixed methods evaluations can highlight the practical challenges that policymakers should bear in mind when designing the content, delivery mode and governance for teacher training programs. For policymakers in Nepal, for example, we highlight: trainers' needs for adequate direction, preparation time and training; the costs to schools of sending teachers to trainings (which may be mitigated by scheduling trainings outside of regular school hours to avoid the need for substitute teachers); the probable need to identify teachers with low subject knowledge and then design differentiated trainings that *teach teachers* "at the right level;" the high time costs to teachers of adopting new daily teaching practices and the small or nonexistent rewards provided by schools that might

compensate them for bearing such costs; and the likely value of devising training content that helps teachers *teach students* (who enter secondary grades with diverse deficits in subject knowledge) at the right level (as suggested by Banerjee et al., 2016, and Glewwe and Muralidharan, 2016).

Our study also contributes to the larger literature on efforts to improve education quality, and to improve the governance of public service provision more generally. Studies by the World Bank (2003) and others on the governance of public service provision have often focused on improving *motivation* by strengthening accountability relationships. Under this conceptualization, this is done by instituting performance pay, undertaking decentralizing or privatizing reforms, or improving the collection and flow of information about performance. Recent reviews of efforts to improve education quality (e.g. Glewwe and Muralidharan 2016, Snilsveit 2015, and World Bank 2018) continue this accountability-focused view of good governance, although they also mention the importance of school management and government "implementation capacity." A separate literature on "management quality" also emphasizes accountability, highlighting the importance of goal setting, performance monitoring and personnel management within schools and other service facilities (Lemos and Scur 2016; Lemos, Muralidharan and Scur 2021).

While our findings are consistent with the importance of ensuring accountability and motivation among frontline service providers, they also suggest an important way in which the current conceptualization of good governance for public services should be expanded. Specifically, good governance involves ensuring not only that agents have adequate *motivation*, but also that they have adequate *time*, *capacity*, *information*, and *budgets* to do their work (Schaffner, 2014, Chapter 13). We showed, for example, that while weak motivation of trainers may have contributed to poor training session implementation, inadequate central-level efforts to identify trainers with sufficient subject expertise and to provide trainers with information, training, preparation time, and training materials seemed likely to be at least as important in diminishing the quality of training sessions. Similarly, while weak teacher motivation may have contributed to low teacher participation in training sessions, the cost to schools of sending teachers (for whom substitutes are difficult to find) to trainings scheduled during regular school days also seems likely to have played an important role. The failure to address in the design of the SSDP program teachers' limited preparation time and limited access to materials also likely exacerbated problems caused by weak accountability. The practical implication is that, in addition to setting clear goals, monitoring performance, and tying rewards to good performance, policymakers and managers must adequately evaluate their agents' initial time, skill, information, and budget constraints. When defining agents' responsibilities, they must then take these constraints into account and either define responsibilities that are feasible given those constraints or provide well-tailored supports (e.g. reduction in other time-consuming responsibilities, remedial training, or additional information or resources) to relax those constraints as necessary to make the more demanding responsibilities feasible.

References

Abeberese, Ama Baafra, Todd Kumler and Leigh Linden. 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

Acharya, S and Uprety, T, 2019. "Evaluating the Design and Impact of the Secondary School Teacher Training Initiative Under the Government of Nepal's School Sector Development Program: A Qualitative Report*."* Unpublished.

Albornoz, Facundo, Maria Anauati, Melina Furman, Mariana Luzuriaga, María Podestá, and Inéz Taylor. 2019. "Training to Teach Science: Experimental Evidence from Argentina", *World Bank Economic Review* 34(2): 393-417.

Araujo, M. Cariddad, Pedro Carneiro, Yyannu Cruz-Aguavo and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten", *Quarterly Journal of Economics,* 131(3): 1415-1453.

Ashraf, Nava, Oriana Bandiera and Kelsey Jack. 2014. "No margin, no mission? A field experiment for public services delivery" *Journal of Public Economics* 120(1): 1-17.

Bando Rosangela, Emma Näslund-Hadley and Paul Gertler 2019. "Effect of Inquiry and Problem-Based Pedagogy on Learning: Evidence from 10 Field Experiments in Four Countries". National Bureau of Economic Research, Working Paper 26280. http://www.nber.org/papers/w26280.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India." No. w22746. National Bureau of Economic Research.

Beg, Sabrin, Adrienne Lucas, Waqas Halim, and Umar Saif. 2019. "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not." Working Paper. National Bureau of Economic Research. https://doi.org/10.3386/w25704. Forthcoming. *American Economic Journal: Economic Policy.*

Beg, Sabrin, Anne Fitzpatrick, and Adrienne M. Lucas. 2021. "Improving Public Sector Service Delivery: The Importance of Management". Department of Economics, University of Delaware.

Benjamini, Yoav and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency". *Annals of Statistics* 29(4): 1165-1188.

Berry, James, Harini Kannan, Shobhini Mukherji and Marc Shotland. 2020. "Failure of frequent assessment: An evaluation of India's continuous and comprehensive evaluation program" *Journal of Development Economics* 143: 102406.

Beteille, Tara, and David Evans. 2019. "Successful Teachers, Successful Students: Recruiting and Supporting Society's Most Crucial Profession". The World Bank. Washington, DC.

Blimpo, Moussa, and Todd Pugatch. 2021. "Entrepreneurship education and teacher training in Rwanda", *Journal of Development Economics* 149: 102583.

Bruns, Barbara, Leandro Costa, and Nina Cunha. 2018. "Through the Looking Glass: Can Classroom Observation and Coaching Improve Teacher Performance in Brazil?" *Economics of Education Review* 64: 214–50. https://doi.org/10.1016/j.econedurev.2018.03.003.

Cardim, Joana, Teresa Molina-Millan and Pedro C. Vicente. 2021. "Can Technology Improve the Classroom Experience in Primary Education? An African Experiment on a Worldwide Program". NOVAFRICA Working Paper Series wp2101, Universidade Nova de Lisboa, Nova School of Business and Economics.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9), 2593-2632.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo and Stephen Taylor. 2020a. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *Journal of Human Resources*, 55(3):926-962.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Mpumi Mohohlwane, and Stephen Taylor. 2020b. "The Challenge of Sustaining Effective Teaching: Spillovers, Fade-out, and the Cost-effectiveness of Teacher Development Programs" gui2de Working Paper, Georgetown University.

Cilliers, Jacobus, Brahm Fleisch, Janeli Kotze, Mpumi Mohohlwane, Stephen Taylor and Tshegofatso Thulare. 2021. "Can Virtual Replace In-person Coaching? Experimental Evidence on Teacher Professional Development and Student Learning in South Africa" RISE Working Paper 20/050.

Clotfelter, Charles, Helen Ladd and Jacob Vigdor. 2010 . "Teacher Credentials and Student Achievement in High School: A Cross Subject Analysis with Fixed Effects." *Journal of Human Resources* 45(3), 655-681.

Damon, Amy, Paul Glewwe, Suzanne Wisniewski and Bixuan Sun, 2019. "What Education Policies and Programmes Affect Learning and Time in School in Developing Countries? A Review of Evaluations from 1990", *Review of Education* 7(2), p295-387.

Das, Jishnu, Stefan Dercon, James Habyarimana and Pramila Krishnan. 2007. "Teacher Shocks and Student Learning. Evidence from Zambia." *Journal of Human Resources* 42(4), 820-862.

de Hoop, Thomas, et al. 2020. "Midline Report for the Mixed-Methods Cluster-Randomized Controlled Trial of Impact Network's eSchool 360 Model in Rural Zambia". American Institutes for Research, Washington, DC.

Dixit, Shanta. 2019. "Seeing through SEE results." Setopati digital newspaper. https://en.setopati.com/view/149213

Duflo, Esther. 2017. "The Economist as Plumber", *American Economic Review* 107(5):1-26.

Duflo Annie, Jessica Kiessel and Adrienne Lucas. 2020. "Experimental Evidence on Alternative Policies to Increase Learning at Scale", National Bureau of Economic Research, working paper 27298. https://doi.org/10.3386/w27298.

Evans, David, and Amina Mendez Acosta. 2021. "Education in Africa: What Are We Learning?", *Journal of African Economies* 30(1):13-54.

Fuje, Habtamu, and Prateek Tandon. 2018. "When do in-service teacher training and books improve student achievement? Experimental evidence from Mongolia". *Review of Development Economics*. 22(3):1360-1380.

Gauri, Varun, Julian Jamison, Nina Mazar and Owen Ozier. 2021 "Motivating bureaucrats through social recognition: External validity—A tale of two states" *Organizational Behavior and Human Decision Processes* 163:117-131.

Gautam, Ganga. 2016. "Teacher Training in Nepal: Issues and Challenges". *Tribhuvan University Journal* 30(2): 43-56. https://doi.org/10.3126/tuj.v30i2.25545.

Glewwe, Paul, and Karthik Muralidharan. 2016 . "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications". In S. Machin, L. Woessmann, and E. A. Hanushek (Eds.), *Handbook of the Economics of Education, Volume 5* (pp. 653-743). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Glewwe, Paul, Celestine Siameh, Bixuan Sun and Suzanne Wisniewski. 2021. "School Resources and Outcomes in Developing Countries", in B. McCall, ed., *The Routledge Handbook of the Economics of Education* Routledge: Abingdon, UK.

Jukes, Matthew, Elizabeth Turner, Margaret Dubeck, Katherine Halliday, Hellen Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski and Simon Brooker. 2017. "Improving literacy instruction in Kenya through teacher professional development and text messages support: A cluster randomized trial". *Journal of Research on Educational Effectiveness*, 10(3):449-481.

Lehrer, Kim, Monica Mawoyo and Samba Mbaye. 2019. "The Impacts of Interactive Smartboards on Learning Achievement in Senegalese Primary Schools". Final Report. International Initiative for Impact Evaluation. New Dehli.

Lemos, Renata, Karthik Muralidharan and Daniela Scur. 2021. "Personnel Management and School Productivity: Evidence from India". RISE Working Paper Series. 21/063. https://doi.org/10.35489/BSG-RISE-WP_2021/063

Lemos, Renata, and Daniela Scur. 2016. "Developing Management: An Expanded Evaluation Tool for Developing Countries." *Manuscript in preparation* 16.007.

Loyalka, Prashant, Anna Popova, Guirong Li and Zhaolei Shi, 2019. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program", *American Economic Journal: Applied Economics* 11(3):128-154.

Lucas, Adrienne, Patrick McEwan, Moses Ngware and Moses Oketch. 2014. "Improving Early Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management*, 33:950–976.

Macdonald, Kevin, Sally Brinkman, Wendy Jarvie, Myrna Machuca-Sierra, Kris McDonall, Souhila Messaoud-Galusi, Siosiana Tapueluelu and Binh Thanh Vu. 2018. "Intervening at Home and Then at School: A Randomized Evaluation of Two Approaches to Improve Early Educational Outcomes in Tonga". World Bank Policy Research Working Paper 8682.

Macdonald, Kevin and Binh Thanh Vu. 2018 "A Randomized Evaluation of a Low-Cost and Highly Scripted Teaching Method to Improve Basic Early Grade Reading Skills in Papua New Guinea". World Bank Policy Research Working Paper 8427.

Ministry of Education. 2016. *School Sector Development Plan: 2016/17–2022/23*. Kathmandu: Ministry of Education, Government of Nepal. https://moe.gov.np/assets/uploads/files/SSDP_Book_English_Final_July_5,_20171.pdf.

Ministry of Education. 2018. *Education in Figures 2017.* Kathmandu: Ministry of Education, Government of Nepal. https://moe.gov.np/assets/uploads/files/Education_in_Figures_2017.pdf

Pant, Bharat Bilas, et al. 2018. " A Study on Identifying Ways for Managing School Level Teachers in Federal System." Dynamic Institute of Research and Development (P) Ltd. Kathmandu. https://www. https://www.doe.gov.np/assets/uploads/files/c3d311abba5aa25f30b60d7cad5bbe0c.pdf.

Pillay, Hitendra, Muttaqi, Iqbal Aziz, Pant, Yagya Raj, & Herath, Nihal. 2017. *Innovative Strategies for Accelerated Human Resource Development in South Asia: Teacher Professional Development - Special Focus on Bangladesh, Nepal, and Sri Lanka.* Asian Development Bank, Philippines.

Piper, Benjamin, Stephanie Simmons Zuilkowski, Margaret Dubeck, Evelyn Jepkemei and Simon J. King 2018. "Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides". *World Development* 106: 324-336.

Piper, Benjamin, and Medina Korda. 2011. "EGRA Plus: Liberia: Program Evaluation Report." RTI International, Research Triangle Park, NC.

Popova, Anna, David Evans, Mary Breeding and Violeta Arancibia, 2019. "Teacher Professional Development around the World: The Gap between Evidence and Practice", Working Paper 517. Washington, DC: Center for Global Development.

Poyck, M. C., et al. "Joint Evaluation of Nepal's School Sector Reform Plan Programme 2009-2016." *Contract* 2014/357774 (2016).

Rauniyar, R, 2019. *Old Fashioned Teacher's Training.* Nagarika Daily.

Republica, 2019. *Public schools fare badly in SEE results.* Retrieved December 11, 2019, from My Republica website: https://myrepublica.nagariknetwork.com/news/68305

Rivkin, Steven, Eric Hanushek and John Kain. 2005. "Teachers, Schools, and Academic Achievement", *Econometrica* 73(2): 417-458.

Schaffner, Julie, 2014. *Development Economics: Theory, Research and Policy Analysis*, John Wiley and Sons, Inc.

Schaffner, Julie, Paul Glewwe and Uttam Sharma. 2018. 'Evaluating the Design and Impact of School Sector Development Program (SSDP) Training for 9th and 10th Grade Math and Science Teachers: Baseline Study Methodology and Findings*',* Unpublished.

Schaffner, Julie, Paul Glewwe and Uttam Sharma. 2019. 'Evaluating the Design and Impact of School Sector Development Program (SSDP) Training for 9th and 10th Grade Math and Science Teachers: Telephone Interview Report', Unpublished.

Schaffner, Julie, Paul Glewwe and Uttam Sharma. 2021. "Evaluation of secondary school teacher training under the School Sector Development Program in Nepal". International Initiative for Impact Evaluation. https://www.3ieimpact.org/sites/default/files/2021-04/GFR-PW3.10-Nepal-SSDP.pdf

Shrestha, Deepika. 2019. 'Evaluating the Design and Impact of School Sector Development Program (SSDP) Training for 9th and 10th Grade Math and Science Teachers: A Report on Training and Video Assignment roll-out (Final Draft)', Unpublished.

Snilsveit, B, Stevenson, J, Phillips, D, Vojtkova, M, Gallagher, E, Schmidt, T, Jobse, H, Geelen, M, Pastorello, M and Eyers, J, 2015. *Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle- Income Countries: A Systematic Review, 3ie Systematic Review 24.* London: International Initiative for Impact Evaluation (3ie).

"World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People*. World Bank. https://openknowledge.worldbank.org/handle/10986/5986 License: CC BY 3.0 IGO."

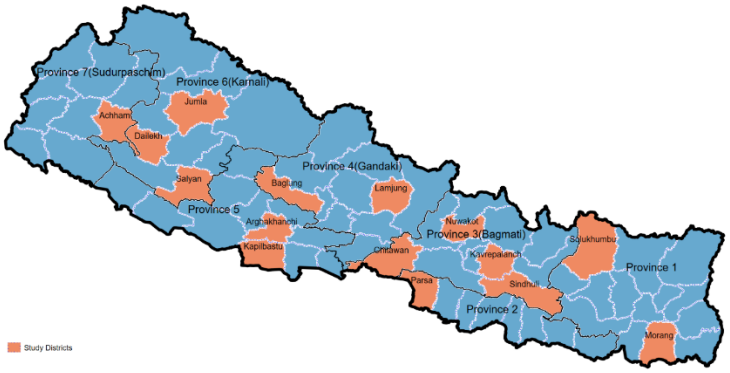
World Bank. 2015. *Conducting classroom observations: analyzing classrooms dynamics and instructional time, using the Stallings 'classroom snapshot' observation system. User guide.* Washington, DC: World Bank.

World Bank. 2018. *World Development Report: Learning to Realize Education's Promise*. The World Bank: Washington DC.

World Bank and Foreign, Commonwealth and Development Office. 2020. "Cost-effective Approaches to Improve Global Learning" The World Bank, Washington, DC, and the UK Foreign, Commonwealth and Development Office, London.

Zhang, Linxiu, Fang Lai, Xiaopeng Pang, Hongmei Yi and Scott Rozelle. 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.

Figure 1. Districts where schools in the study are located



Note: This map was generated using open access files at https://opennepal.wordpress.com/ and https://gadm.org/data.html.

Table 1. Description of Student Attrition Among Students in Grades 8 and 9 at Baseline

	Students in Grade 8 Baseline			Students in Grade 9 Baseline		
	Treatment Schools	Control Schools	P-value for test of equality[a]	Grade 9 Treatment Schools	Grade 9 Control Schools	P-value for test of equality[a]
Percent of baseline students who were not tested at endline (i.e. attriters), of which:	35.4	35.9	0.867	44.3	41.4	0.065
Enrolled in grade but absent	21.2	17.8		27.1	25.1	
Not in school	7.9	9.0		7.2	7.2	
Classes not in session	0.0	0.0		6.3	5.9	
Other reasons (repeating previous grade, moved to another school, or unknown)	6.3	9.1		3.7	3.3	
Number of students	3743	3906		4167	4614	
Average baseline math test scores[b] among those:						
Tested at endline	0.01	0.01	.645	0.08	0.21	.232
Not tested at endline (attriters)	-0.33	-0.21	.063	-0.20	-0.19	.235
Average baseline science test scores[b] among those:						
Tested at endline	0.05	0.03	.826	0.08	0.17	.613
Not tested at endline (attriters)	-0.41	-0.30	.142	-0.25	-0.20	.160

Notes: [a]Tests of hypothesis that coefficient on "treat" indicator is zero in weighted student-level regressions of the dependent variable on treat and district-stratum fixed effects, with standard errors adjusted for stratification and school-level clustering. For attrition rate tests, the sample includes all students tested at baseline, and the dependent variable is an indicator of attrition (i.e. not being tested at endline). For average test score comparisons of those tested (not tested), the sample is all baseline students who were (were not) tested at endline, and the dependent variable is the test score. [b]Test scores are indices constructed from joint IRT analysis of baseline and endline scores for a given grade-level cohort and subject.

Table 2: Baseline descriptive statistics and balance tests: schools and teachers

Variable	Number of obser-vations	Mean for Treated Sample (std. dev.)	Mean for Control Sample (std. dev.)	p-value for test of $\beta_T = 0$[a,b,c,d]
School-level characteristics				
Total number of students in school	203	402.1 (214.5)	453.1 (277.7)	0.036**
Hours walking to nearest all-weather road	203	2.85 (3.64)	3.47 (5.26)	0.395
Students per section in Grade 9	203	47.93 (25.47)	51.11 (27.82)	0.242
Students per section in Grade 10	203	41.62 (20.73)	44.20 (23.19)	0.279
Days school was open last year (Grade 9)	203	194.87 (18.36)	196.31 (13.68)	0.544
School has electricity (several hours most days)	203	0.81 (0.40)	0.73 (0.44)	0.085*
Whether head teacher has at least Master's degree	203	0.63 (0.49)	0.52 (0.50)	0.078*
Hours per week head teacher teaches	203	16.95 (6.49)	15.81 (7.08)	0.193
Estimated management quality index	201	0.04 (0.90)	-0.07 (0.89)	0.454
Teacher-level characteristics				
Is female	395	0.09 (0.29)	0.04 (0.20)	0.030**
Has at least Bachelor's degree in math/science	361	0.79 (0.41)	0.82 (0.38)	0.291
Had SSRP training	395	0.30 (0.46)	0.32 (0.47)	0.481
Years of experience	393	10.69 (7.80)	11.49 (8.42)	0.168
Hours spent preparing for class	393	0.79 (0.85)	0.82 (1.07)	0.505

Notes: [a] For all p-values, * indicates significance at the 10% level, ** indicates significance at the 5% level, and *** indicates significance at the 1% level. [b]The p-values from tests of the hypothesis that the coefficient on Treat is zero, based on WLS regressions of each variable on a treatment indicator and district and priority stratum fixed effects. [c]For binary outcome variables, weighted probit regressions were used instead of WLS. [d] All p-values are greater than 0.10 after adjustment for multiple hypothesis testing.

Table 3: Baseline descriptive statistics and balance tests: students

Variable	Number of observa-tions	Mean for Treated Sample (std. dev.)	Mean for Control Sample (std. dev.)	p-value for test of $\beta_T = 0$[a.b.c,d]
Is female	16435	0.55 (0.50)	0.55 (0.50)	0.875
Father can read and write	15594	0.83 (0.37)	0.82 (0.38)	0.470
Father has at least secondary education	15753	0.28 (0.45)	0.28 (0.45)	0.881
Mother can read and write	14830	0.59 (0.49)	0.59 (0.49)	0.964
Mother has at least secondary education	15831	0.11 (0.31)	0.12 (0.33)	0.384
Nepalese is main language spoken at home	16251	0.74 (0.44)	0.77 (0.43)	0.226
Family IRT asset index[d]	16435	-0.03 (0.77)	-0.04 (0.77)	0.950
Grade 8 math percentage score	7651	18.13 (10.52)	18.11 (10.62)	0.460
Grade 8 math IRT latent variable	7651	0.01 (0.86)	0.01 (0.87)	0.414
Grade 8 science percentage score	7651	28.58 (13.07)	27.76 (11.66)	0.609
Grade 8 science IRT latent variable	7651	0.08 (0.95)	0.03 (0.86)	0.477
Grade 9 math percentage score	8784	27.95 (15.00)	29.06 (16.18)	0.168
Grade 9 math IRT latent variable	8784	-0.01 (0.88)	0.04 (0.95)	0.196
Grade 9 science percentage score	8784	27.61 (10.61)	28.48 (11.54)	0.314
Grade 9 science IRT latent variable	8784	-0.01 (0.81)	0.04 (0.88)	0.366

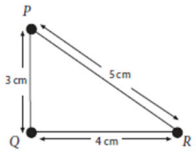
Notes: [a] For all p-values, * indicates significance at the 10% level, ** indicates significance at the 5% level, and *** indicates significance at the 1% level. [b]The p-values from tests of hypothesis that coefficient on a treatment indicator is zero, based on WLS regressions of each variable on the treatment indicator and district and priority stratum fixed effects. [c]For binary outcome variables, weighted probit regressions were used instead of WLS. [d]The family asset IRT index is defined in Section 3.4. [d] All p-values are greater than 0.10 after adjustment for multiple hypothesis testing.

Table 4: ITT estimates of impact of SSDP training on students' normalized test Scores, full endline and panel samples

	Full Sample Weighted Mean (Std. Dev.)		Full Sample Estimates	Sample Size	Panel Sample Estimates		Sample Size
	Treated Schools	Control Schools	Treat[a,b,c,d]		Treat[a,b,c,d]	Baseline Test Score[b,c]	
Full assessments							
Grade 9 math	-0.057 (0.931)	0.000 (1.000)	-0.110* (0.066)	6,800	-0.107** (0.050)	0.532*** (0.020)	4,903
Grade 9 science	-0.051 (0.912)	0.000 (1.000)	-0.109* (0.060)	6,797	-0.106* (0.054)	0.494*** (0.022)	4,901
Grade 10 math	-0.035 (0.998)	0.000 (1.000)	-0.044 (0.072)	5,832	-0.000 (0.050)	0.563*** (0.017)	4,992
Grade 10 science	-0.002 (0.971)	0.000 (1.000)	0.006 (0.074)	5,829	0.025 (0.061)	0.502*** (0.021)	4,990
SSDP focus items							
Grade 9 math	-0.004 (0.953)	0.000 (1.000)	-0.046 (0.066)	6,800	-0.054 (0.056)	0.444*** (0.021)	4,903
Grade 9 science	-0.061 (0.914)	0.000 (1.000)	-0.100* (0.057)	6,797	-0.075 (0.054)	0.426*** (0.020)	4,901
Grade 10 math	-0.044 (1.001)	0.000 (1.000)	-0.037 (0.070)	5,832	-0.024 (0.059)	0.444*** (0.018)	4,992
Grade 10 science	0.009 (0.979)	0.000 (1.000)	0.024 (0.072)	5,829	0.022 (0.063)	0.433*** (0.019)	4,990

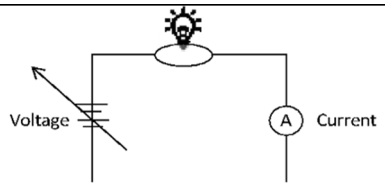
Notes: [a]Estimates of β_T from WLS regressions of normalized student assessment scores on the treat variable, district by priority stratum fixed effects, and dummy variables for whether assent was requested before or after the test and whether the math test was given first (followed by the science test). Panel estimates add baseline scores. [b]Standard errors, in parentheses, account for random assignment within strata and are clustered at the school level. [c]Statistical significance at .10, .05 and .01 levels indicated by *, ** and ***. [d]All p-values are greater than 0.10 after adjustment for multiple hypothesis testing.

Table 5: Math Teacher Responses to Evaluations of Student Assessment Items[a]

Assessment item	Percent selecting correct answer[b]
1. Which of these has the same value as 342? a. 3000 + 400 + 2 c. 30 + 4 + 2 b. 300 + 40 + 2 d. 3 + 4 + 2	97.2
2. All of the students in a class cut out paper shapes. The teacher picked one out and said "this shape is a triangle." Which of the following statements MUST be correct? a. The shape has three sides c. The shape has equal sides b. The shape has a right angle d. The shape has equal angles	92.7
3. It takes Diksha 4 minutes to wash a window. She wants to know how many minutes it will take her to wash 8 windows at this rate. She should: a. Multiply 4 by 8 c. Subtract 4 from 8 b. Divide 8 by 4 d. Add 8 and 4	97.6
4. Which of the numbers below is equal to $\frac{7}{10}$? a. 70 b. 7 c. 0.7 d. 0.07	91.5
5. There were *m* boys and *n* girls in a parade. Each person carried 2 balloons. Which of these expressions represents the total number of balloons that were carried in the parade? a. 2(*m* + *n*) c. 2*m* + *n* b. 2 + (*m* + *n*) d. *m* + 2*n*	96.3
6. What is 15 × 9? a. 100 b. 135 c. 130 d. 531	91.9
7. Which of these is the reason that triangle *PQR* is a right-angle triangle?  a. $3^2 + 4^2 = 5^2$ b. $5 < 3 + 4$ c. $3 + 4 = 12 - 5$ d. $3 > 5 - 4$	98.8
8. A thin wire 20 centimeters long is formed into a rectangle. If the width of this rectangle is 4 centimeters, what is its length? a. 5 centimeters c. 12 centimeters b. 6 centimeters d. 16 centimeters	60.6
9. If x + 3y = 11 and 2x + 3y = 13, then y = ? a. 3 b. 2 c. -2 d. -3	80.1
10. If the volume of a cube is 216 cubic cm, what will be the side of the cube? a. 36 cm b. 6 cm c. 54 cm d. 24 cm	91.5
11. What is the sum of mode and median of the following data? 12, 15, 11, 13, 18, 11, 13, 12, 13 a. 26 b. 31 c. 36 d. 25	83.7

Notes: [a]Sample = 246 teachers. The percentages reported do not employ population weights. They are intended to describe only the sample of teachers who completed the forms voluntarily and anonymously. [b] Based on response to the question: "Which answer do you think the item's designer had in mind as the correct answer?"

Table 6: Science teacher responses to evaluations of student assessment items[a]

Assessment Item	Percent correctly answered[b]
1. All living things can be grouped as plants or animals. Which of these in the list below are ANIMALS? Fish Fern Man Grass Algae Crocodile a. All are animals c. Algae, Fern and crocodile are animals b. All are plants d. Fish, man, and crocodile are animals	93.2
2. What is the chemical formula of water? a. H_20 b. NaCl c. NaOH d. H_2O_2	97.4
3. What is the rate of change in velocity per unit of time? a. Acceleration b. Relative velocity c. Speed d. Velocity	92.7
4. What is the main function of red blood cells? a. To fight disease in the body c. To remove carbon monoxide from all parts of the body b. To carry oxygen to all parts of the body d. To produce blood proteins which cause blood to clot	79.1
5. Which of the following is the major cause of tides? a. Evaporating ocean water by the heat of the sun c. Earthquakes on the ocean floor b. Gravitational pull of the moon d. Changes in wind direction	85.0
6. Which one of the following statements about liquid evaporation is correct? When a liquid evaporates: a. The temperature in the air above the liquid decreases b. Fast-moving liquid molecules near the surface escape to the air and the liquid gets warmer c. The gas pressure of the substance directly above the liquid depends only on atmospheric pressure d. Fast-moving liquid molecules near the surface escape to the air and the liquid gets colder	34.6
7. Which of the following grows from a seed? a. Ant b. Grass c. Mosquito d. Caterpillar	88.0
8. When a small volume of water is boiled, a large volume of steam is produced. Why? a. The molecules are further apart in steam than in water b. Water molecules expand when heated c. The change from water to steam causes the number of molecules to increase d. Atmospheric pressure works more on water molecules than on steam molecules	52.4
9. Some students used an ammeter *A* to measure the current in the circuit for different voltages. The table below shows some results. Voltage (volts) | Current (milliamperes) 1.5 | 10 33.0 | 20 6.0 |  What is the missing value? a. 30 b. 40 c. 50 d. 60	75.6
10. Which one of the following statements best describes a comet? a. A comet is made of an icy substance and dust particles b. A comet is smaller than the sun c. A comet is very close to the sun d. A comet revolves around the sun in highly elliptical orbit and is made up of an icy substance	59.8
11. The symbol of the element nitrogen is: a. N b. He c. O d. H	96.2

Notes: [a]Sample size = 234 (except 233 for question 8) The percentages reported do not employ population weights. They are intended to describe only the sample of teachers who completed the forms voluntarily and anonymously. [b] Based on response to the question: "Which answer do you think the item's designer had in mind as the correct answer?"

Table 7: Summary of ITT estimates of impact on teacher subject knowledge and teacher attendance

Intermediate Outcomes	Sample Size	Mean (std. dev. in parentheses)		p-value of test of no impact[b]
		Control	Treatment	
Math knowledge[a]	246	0.000 (1.000)	-0.014 (1.010)	0.907
Science knowledge[a]	233	0.000 (1.000)	-0.136 (0.966)	0.298
Teacher present on 1st day of school visit (enumerator observation) (%)	434	91.5 (27.9)	90.5 (29.32)	0.422

Notes: [a]Test scores are normalized by subtracting the mean, and dividing by the standard deviation, of the control group. [b]Teacher subject knowledge is based on weighted least squares regression; teacher presence is based on a probit regression.

Table 8: ITT estimates of impacts on teacher teaching practices (student reports)[a]

	Mathematics			Science		
	Descriptive Statistics		Test of No Impact	Descriptive Statistics		Test of No Impact
	Control	Treatment	(p-value)[b]	Control	Treatment	(p-value)[b]
Gives homework frequency (% distribution)	--	--	0.960	--	--	0.793
Up to once a week	} 19.6	} 19.4	--	14.9	16.0	--
2-3 times a week			--	33.0	32.0	--
Every day	80.4	80.6	--	52.1	52.0	--
Homework checking frequency (% distribution):	--	--	0.522	--	--	0.639
Never	} 14.8	} 15.4	--	4.3	3.7	--
Up to once a week			--	19.7	22.3	--
2-3 times a week	32.1	33.9	--	35.7	36.2	--
Every day	53.1	50.6	--	40.3	37.8	--
Homework correction frequency (% distribution):	--	--	0.925	--	--	0.453
Never	} 13.9	} 15.1	--	4.6	4.9	--
Up to once a week			--	16.1	17.7	--
2-3 times a week	27.9	26.1	--	33.7	33.5	--
Every day	58.2	58.8	--	45.6	43.9	--
Interactive teaching frequency (% distribution):	--	--	0.272	--	--	0.519
Up to once a week	14.0	16.8	--	18.1	18.7	--
2-3 times a week	32.7	32.7	--	35.9	37.2	--
Every day	53.3	50.5	--	46.0	44.1	--
Class time group work frequency (% distribution)	--	--	0.755	--	--	0.985
Never	32.5	31.4	--	26.9	25.7	--
Less than once a week	7.6	6.8	--	7.4	7.5	--
Once a week	16.9	20.5	--	19.1	21.7	--
2-3 times a week	26.6	27.8	--	28.3	29.5	--
Every day	16.4	13.4	--	18.2	15.7	--
Frequency of using local materials or visual aids (% distribution):	--	--	0.709	--	--	0.951
Never	26.0	24.1	--	18.7	17.0	--
Less than once a week	11.9	11.5	--	10.1	10.7	--
Once a week	18.4	24.4	--	21.8	24.8	--
2-3 times a week	24.6	24.1	--	29.3	29.8	--
Every day	19.2	15.9	--	20.1	17.8	--
Frequency of using materials from internet (% distribution):	--	--	0.904	--	--	0.391
Never	32.2	33.4	--	23.5	25.3	--
Less than once a week	9.9	9.4	--	9.9	10.7	--
Once a week	18.4	21.4	--	20.4	22.1	--
2-3 times a week	23.3	23.0	--	29.6	28.6	--
Every day	16.2	12.8	--	16.5	13.3	--

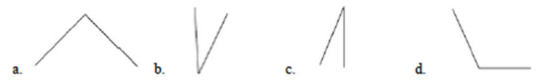
[a]Sample size ranges from 12,482 to 12,574. The estimate for "gives homework all days" is a probit; all other estimates are ordered probits. For dichotomous outcomes, the descriptive statistics columns report the weighted percentage for which the outcome is true. For ordered polychotomous variables, they report the weighted percentage distributions by category. [b] From regressions of the dependent variable on the treatment indicator and strata dummy variables, using weighted probit estimation for binary variables and weighted ordered probit estimation for ordered categorical variables. The test statistics account for the stratified sample design and clustered standard errors at the school level.

Table 9: ITT Estimates of impacts on teaching practices: (head teacher reports)[a]

Intermediate outcome	Sample Size	Descriptive Statistics		Estimation Method for Testing	Test of No Impact (p-value)[b]
		Control	Treatment		
Teacher ever creates teaching materials from local resources (%)	437	39	41	Probit	0.96
Teacher's frequency of using teaching materials or visual aids (%distribution)	438	--	--	Ordered Probit	0.072*
Never		22	17	--	--
Sometimes (less than once per week)		70	65	--	--
Often (one or more times per week)		9	18	--	--
Teacher ever collects or requires students to collect local information (%)	422	65	65	Probit	0.816
Teacher frequency requiring students to work in small groups (% distribution)	441	--	--	Ordered Probit	0.079*
Never		20	14	--	--
Sometimes (less than once per week)		67	67	--	--
Often (one or more times per week)		13	13	--	--
Teacher ever requires students to work on longer term projects (%)	439	51	59	Probit	0.013**

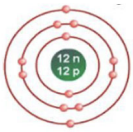
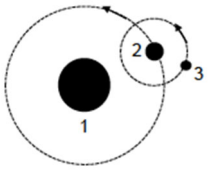
[a]The unit of observation is the teacher. [b]From regressions of the dependent variable on the treatment indicator and strata dummy variables, using weighted probit estimation for binary variables and weighted ordered probit estimation for ordered categorical variables. Test statistics account for the stratified sample design and clustered standard errors at the school level.

Table 10: Student performance on below grade-level math questions

Question	Grade level[a]	9th Graders		10th Graders	
		Number Tested[b]	Percent Correct[c]	Number Tested[b]	Percent Correct[c]
1. If A={a, e, i, o, u}, what is the value of n(A)? a. 4 b. 5 c. 3 d. 2	8 (BL)	6801	74.7	5833	81.7
2. If a square has a side of 6 cm, what is its area? a. 24 cm^2 b. 36 cm^2 c. 12 cm^2 d. 64 cm^2	8 (BL)	3404	61.7	2899	67.1
3. What is the volume of a cube with a side of 2 cm? a. 16 cm^3 b. 4 cm^3 c. 6 cm^3 d. 8 cm^3	8 (BL)	6801	44.0	5833	54.3
4. Which of the following is NOT a parallelogram? a. Rectangle c. Rhombus b. Square d. Trapezoid	8 & 9 (BL)	3404	25.1	2899	28.9
5. What is 15 x 9? a. 100 b. 135 c. 130 d. 531	7 (YL)	3397	86.8	2934	91.1
6. What is 6 ÷ 3? a. 18 b. 2 c. 3 d. 9	7 (YL)	3404	79.3	2899	86.7
7. One of these angles is a right angle. Which one?  	4 (TIMSS)	3397	46.0	2934	51.9
8. Shaheen has 2 pencil boxes. Each box has 5 pencils. How will you find the total number of pencils in the two pencil boxes? a. 2 + 5 b. 5 - 2 c. 2 x 5 e d. 2 + 2	4 (SLS)	6801	66.9	5833	71.5

[a]BL indicates an assessment item from the baseline assessment. YL is an assessment item from the Young Lives study (see www.younglive.org.uk). TIMSS is an assessment item from the Trends in International Mathematics and Science Study (see e). SLS is an assessment item from the Student Learning Study conducted in India (see www.ei-india.com/study_on_student_learning). [b]The number of students varies by question because some questions were used on both versions of the tests while others were used only on one version[c] The percent correct are unweighted averages.

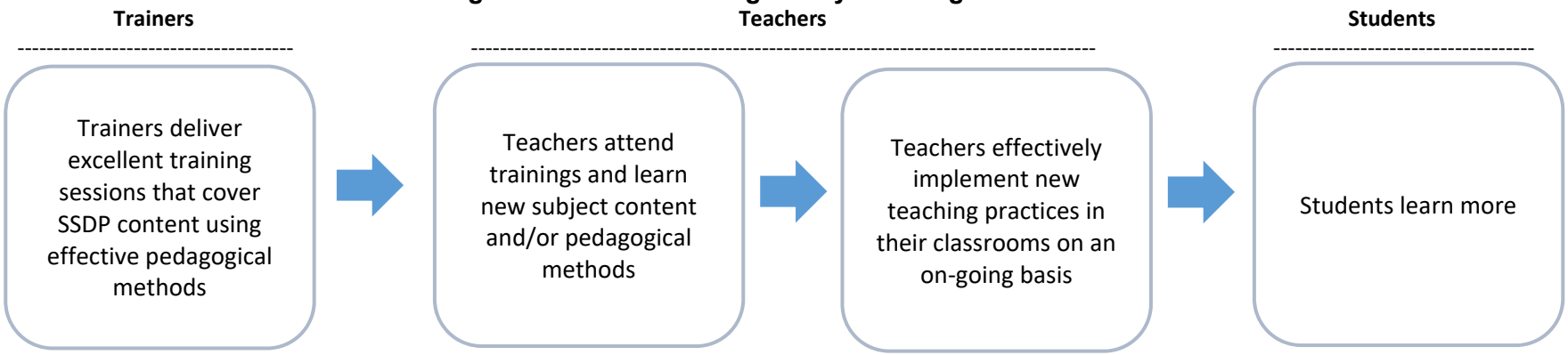
Table 11: Student performance on below grade level-science questions

Questionee	Grade level[a]	9th Graders		10th Graders	
		Number Tested[b]	Percent Correct[c]	Number Tested[b]	Percent Correct[c]
1. What is the rate of change in velocity per unit of time? a. Acceleration b. Relative velocity c. Speed d. Velocity	8 (BL)	6801	74.9	5833	70.7
2. What is the name of the liquid used in the instrument below? a. Water b. Alcohol c. Mercury d. Milk	8 (BL)	6801	65.7	5833	68.7
3. Study the given atomic structure and select the name of the element?  a. Calcium b. Oxygen c. Magnesium d. Sodium	8 (BL)	3368	62.3	2883	66.7
4. Which of the following is a satellite of a planet? a. Earth b. Mercury c. Jupiter d. Moon	8 (BL)	6801	52.3	5833	53.6
5. Which of the following grows from a seed? a. Ant b. Grass c. Mosquito d. Caterpillar	4 (QES)	3433	81.6	2950	78.8
6. Mahesh gave some good reasons why kettles and kitchen pans are often made of copper. Which reason is correct? a. Copper is a good conductor of heat. b. Copper is easy to melt c. Copper is difficult to shape d. Copper dissolves in hot water	4 (TIMSS)	3433	67.0	2950	74.6
7. The figure shows Earth, the Moon, and the Sun. Each body is labeled by a number. The arrows show the direction each body is moving. Which is the body labeled 2?  a. The Earth b. The Moon c. The Sun d. It is not possible to say with the information provided	4 (TIMSS)	3433	49.8	2950	55.1
8. Neeraj put a thermometer in a glass filled with hot water. Why does the liquid inside the thermometer rise? a. Gravity pushes it up b. Air bubbles are released. c. Heat from the water makes it expand d. Air pressure above the water pulls it up	4 (TIMSS)	3368	34.6	2883	40.6

Notes: [a]BL indicates assessment items from the baseline assessment. TIMSS is an assessment item from the Trends in International Mathematics and Science Study (see www.timssandpirls.bc.edu). QES is an assessment item from the Quality Education Study in India (see e). [b]The number of students for each question varies because some questions were used on both versions of the tests while others were used only on one version.
[c]The percent correct are unweighted averages.

Appendix A

Figure A1. SSDP training theory of change



Assumptions:

Capacity	- Trainers have adequate command of math and science subject content - Trainers have adequate guidelines, training and skill for translating SSDP outline into 10 days of high-quality training plans and materials	- Teachers have permission from their schools to attend - Teachers have adequate command of math and science subject content to understand SSDP content - Teachers do not already have the knowledge and skills covered in SSDP trainings	- Teachers have adequate skill to translate ideas learned at trainings into new lesson plans for most class sessions - Teachers are willing to experiment with new teaching methods, which may be difficult to execute at first and may not be well received by students	- Students have adequate background knowledge acquired in previous grades to understand grade 9 and 10 content, and thus to benefit from improved methods for teaching this material
Resources	- Trainers have adequate teaching supplies and access to adequate facilities - Trainers have adequate time to prepare the training plans and materials	- Teachers are provided with adequate per diems and/or room and board	- Teachers have time to devote to preparation of teaching materials for demonstration-based methods - Teachers have adequate means for acquiring necessary teaching materials	- Students receive adequate nutrition and rest at home that they can concentrate in school and thus benefit from improved teaching methods
Motivation	- Trainers are well motivated to provide transformative trainings	- Trainers perceive personal or professional benefits to attendance that outweigh the costs	- Teachers perceive personal or professional benefit to creating new lesson plans and materials for most class sessions that outweigh the costs	- Students are motivated to pay attention and study, and thereby benefit from improved teaching methods

Appendix B. Review of Studies Estimating Student Learning Impacts for Teacher Training Programs in Low- and Middle-Income Countries

To situate our study in the relevant teacher training literature, we reviewed all rigorous evaluations of teacher training programs in low- and middle-income countries that have been published in peer-reviewed journals since 2000 or have appeared as working papers in the last five years (since 2016). This is difficult to do systematically, since "teacher training" takes many forms and is often bundled with other efforts to improve teaching and learning. We define teacher training broadly, including in-person and virtual coaching and mentoring, so we searched for evaluations where such training was a major component or the sole component. We include only interventions intended to improve teaching during the regular school day at the primary or secondary level, and only "high quality" studies, defined as those that: (a) seek to identify causal impacts on student learning using randomized control trials, difference-in-differences estimation, or regression discontinuity; (b) cluster standard errors appropriately; and (c) provide adequate information on program design, sampling, estimation, and standard errors. We began by selecting relevant references in Glewwe et al. (2021), Evans and Mendez Acosta (2021) and Popova, et al. (2019). We then used Econlit and Google Scholar searches to find more recent publications or working papers. To avoid double counting, we excluded papers analyzing midline results when papers analyzing endline outcomes were available.

Table B1 summarizes the features of the interventions, evaluation methods, and results for the 23 papers identified by our search. For each paper, we provide details on up to two interventions: the primary intervention and (if relevant) an alternative intervention. (Some papers report on additional interventions not included in our table, either because they studied more than two teacher training intervention variants or because they also studied interventions that did not include teacher training.) When selecting "main" effect size estimates (in standard deviations of the baseline control test score distribution) to report, we choose (as relevant) estimates for the longest period of exposure, for aggregate scores rather than sub-scores, and for reading comprehension rather than other language scores. Most are estimates of ITT impacts.

The superscripts SC and LF indicate (separately for primary and alternate interventions) whether the interventions included design features that could be labeled structured curriculum support and long-term follow-up. We defined structured curriculum support broadly to include provision of lesson plans, teaching aids, worksheets, or other supports to facilitate teachers' application of new teaching approaches (to which they were exposed in the trainings) into their daily classes, for all the topics they must cover in the school year. (We also include an intervention that provided teachers videos of expert teachers teaching concepts throughout the curriculum, which they could use either to obtain examples or to show in the classroom, because, as with provision of lesson plans, this intervention may streamline teachers' daily preparation.) We defined longer-term follow-up to include periodic monitoring and coaching (or multiple training sessions over one or more school years) through which teachers are observed after initial training.

Table B1. Summary of High-Quality Evaluations of Teacher Training Interventions in Developing Countries[a]

Study	*Country*	*Grades*	*Primary Intervention Design*			*Who Implemented the Training?*	*Scale of Program/ Number of Schools*	*Method*	*Alternative intervention evaluated*	*Main Impact Estimates (Std. Err.)[b]*		
			Length of Group Trainings	*Follow-up*	*Other Support or Services Provided*					*Subject*	*Primary Intervention*	*Alternative Intervention*
Abeberese et al. 2014	Philip-pines	4	2 days	Monitoring, unspecified support[LF]	Classroom library	NGO	Province /100	RCT	None	Reading	0.06 (0.03)	Not applicable
Albornez et al. 2020	Argentina	7	None[c]	Weekly coaching[LF]	Teaching guides, student materials[SC]	University	City /70	RCT	Without coaching[SC]	Science	0.65 (0.14)	0.55 (0.13)
Bando et al. 2019	Argentina, Belize, Peru	3 or 4	29-73 hours	Tutoring or mentoring[LF]	Lesson plans, student materials for problem-based pedagogy[SC]	Not stated	2-3 countries /280 for science, 399 for math	RCT	None	Math (Arg.) Math (Belize) Sci. (Arg.) Sci. (Belize) Sci. (Peru 2010) Sci. (Peru 2012)	0.13 (0.06) 0.16 (0.09) 0.08 (0.04) 0.29 (0.09) 0.17 (0.08) 0.14 (0.08)	Not Applicable
Beg et al. 2019	Pakistan	8	2 days	None	Teacher tablet, LED screen, projector [d,SC]	Government and university	3 districts /130	RCT	None relevant	Math/Science	0.30 (0.13)	
Beg et al. 2021	Ghana	4-6	10 days divided across 3 terms	None specified	Teacher guides and materials for differentiated instruction[SC]	Government	20 districts /210	RCT	Add training for principals and supervisors[SC]	Math English	0.140 (0.026) 0.065 (0.022)	0.131 (0.029) 0.076 (0.024)
Berry et al. 2020	India	2-5, 8	7 days	Monthly monitoring[LF]	Materials for continuous evaluation	Companies contracted by government	2 districts /500	RCT	Training for teaching at right level (Gr 2-5)[LF,SC]	Grades 2-5 Hindi math Grade 8 Hindi Grade 8 math	 0.028 (0.021) 0.014 (0.021) 0.022 (0.045) 0.040 (0.056)	 0.135 (0.021) 0.023 (0.022) - - - -
Blimpo and Pugatch 2021	Rwanda	10-12	Six 4-day trainings over 2 years	6+ youth leader visits	School business club promotion	NGO	3 provinces /207	RCT	None	Entrepreneur-ship	-0.01 (0.05)	Not applicable
Bruns et al. 2018	Brazil	10-12	None	Coaching by remotely trained coach[LF]	Pedagogy & class-mgmt. book, video examples	Company	1 state /292	RCT	None	Math Portuguese	0.081 (0.032) 0.055 (0.011)	Not applicable
Cardim et al. 2021	Angola	4-6	20 hours	Weekly trainer visits[LF]	CAL software, teacher computer, student tablets	NGO	City/42	RCT	None	Math Portuguese Science	-0.015 (.061) -0.007 (.061) 0.066 (.028)	Not applicable

Cilliers et al. 2020a	South Africa	1-2	Two 2-day trainings per year	~3 trainer visits[LF]	Lesson plans, learnineg materials[SC]	NGO	2 districts /180	RCT	Add monthly coaching[LF,SC]	Reading	0.177 (0.081)	0.290 (0.080)
Cilliers et al. 2021	South Africa	1-3	Two 2-day trainings per year	In-person monthly coaching[LF]	Lesson plans, learning materials[SC]	NGO	2 districts /180	RCT	Virtual coaching[LF,SC]	Engl. reading Engl. oral Local lang.	0.130 (0.068) 0.313 (0.068) -0.047 (0.068)	-0.047 (0.069) 0.123 (0.072) -0.193 (0.074)
De hoop et al. 2020	Zambia	1	Weekly training	None (but training is weekly)[LF]	Solar power, projector, tablets with e-learning curriculum, lesson plans[SC]	NGO	3 districts /63	RCT	None	Reading Math	0.404 (0.083)[e] 0.219 (0.065)[e]	Not applicable
Duflo et al. 2021	Ghana	1-3	Length not specified	None specified	Differentiated instruction aids[SC]	Government	National/ 200	RCT	None relevant	English Math Local lang.	0.085 (0.047) 0.056 (0.043) 0.060 (0.061)	Not applicable
Fuje, Tandon 2018	Mongolia	3-4	3 days	Mentoring visit[LF]	Classroom library	Government	National /172	RCT with PSM	None relevant	Reading Writing Math	0.257 [0.00] 0.271 [0.02] 0.259 [0.00]	
Jukes et al. 2017	Kenya	1-2	3 days +annual refresher	Two-way SMS support	Semi-scripted lesson plans[SC]	Researchers	2 districts /101	RCT	None	English Swahili Numeracy	0.12 (0.065) 0.13 (0.065) -0.15 (0.075)	Not applicable
Lehrer et al. 2019	Senegal	1-2	2-5 days	Technical support	Interactive whiteboard, digital content, computer, projector	Government	National /122	DD	None	French Math Soc. Sci.	.109 (.095) .186 (.090) .185(.115)	
Loyalka et al. 2019e	China	7-9	15 days	None	Online resources and communication	Government	94 counties (1 province) /300	RCT	Add SMS and phone follow-up[f]	Math	-0.006 (.034)	0.005 (0.035)
Lucas et al. 2014	Kenya, Uganda	1-3	12 days	Monthly mentoring[LF]	Libraries	NGO	4 districts /221	RCT	None	Written literacy Swahili-Kenya Lango-Uganda	 0.024 (0.032) 0.199 (.045)	Not applicable
Macdonald et al. 2018	Tonga	1-2	Not stated	Unspecified monitoring, coaching[LF]	Scripted lesson plans, materials[SC]	Government	National /73	RCT	None	Reading	0.33 (0.06)	Not applicable
Macdonald and Vu 2018	Papua New Guinea	3-4	Not stated	Unspecified mentoring, coaching[LF]	Highly scripted lesson plans[SC]	Government	2 provinces /68	RCT	None	Reading	0.456 (0.239)	Not applicable

Piper and Korda, 2010	Liberia	2-3	1 week per semester	Monthly coaching[LF]	Tightly scripted lesson plans, assessments, community engagement[SC]	NGO with government support	National /177	RCT with DD	Only community engagement	Reading	0.82 (0.07)	0.02 (0.07)
Piper et al. 2018	Kenya	1-2	10 days per year	Coaching by trained curriculum support officers[LF]	None	Government	2 counties /171	RCT with DD	Add text-books, plus partially scripted lesson plans[LF,SC]	Kiswahili English Math	-0.11 (0.31) 0.41 (0.23) 0.34 (0.18)	1.14 (0.219) .91 (0.297) 0.16 (0.236)
Zhang et al. 2013	China	4-5	3 weeks	None specified	None specified	Researchers	City /70	RCT	None	English	0.12 (0.21)	Not applicable

Notes:

[a] We defined "high quality studies" as those that use randomized control trial (RCT), difference-in-differences (DD), or regression discontinuity design (RDD) methods, and use clustered standard errors as appropriate.

[b] When choosing main impact estimates from multiple reported impacts, we select (as relevant) estimates for: the longest period of exposure; aggregate scores rather than sub-component scores; and reading comprehension rather than other language scores. Most estimates are ITT effects. In some cases, standard errors are approximated from confidence interval reports. Numbers in square brackets are p-values, which we report when standard errors are not reported.

[c] All teachers in treatment and control groups had taken standard 4-hour government in-service training.

[d] Tablets contained short videos of expert teachers explaining content, electronic textbook, review questions and simulations of complex ideas such as photosynthesis.

[e] These are estimates of ITT impact among children age 6-9 near school who are eligible for grade 1, rather than among students. Only a little more than half of sampled children are in school, so estimated TOT impacts, which include effect of some increased enrollment, are twice as large as these reported ITT impacts.

[f] The study included an additional study arm that added to the main training an anticipated evaluation of teaching practice two months after training. No effect was detected for this intervention.

[SC] Included in our category of programs providing some form of "structured curriculum support."

[LF] Included in our category of programs providing some form of longer-term follow-up (after the initial group training).