Evaluation of the design, implementation and impacts of School Sector Development Program trainings for teachers of secondary math and science in Nepal

Julie Schaffner*
The Fletcher School, Tufts University
Julie.Schaffner@tufts.edu

Paul Glewwe
Department of Applied Economics, University of Minnesota

Uttam Sharma
Independent Consultant

Impact Evaluation Report
February 2020 [Updated November 2020]

Acknowledgements

We gratefully acknowledge the International Initiative for Impact Evaluation (3ie) for funding this evaluation under Grant PW3.10.NP.IE. We thank the Government of Nepal's National Planning Commission for leading this research, and thank the Government of Nepal's Ministry of Education, Science and Technology, Center for Education and Human Resource Development, and especially the former Department of Education and former National Center for Educational Development (which have been merged into the Center for Education and Human Resource Development), for their collaboration. We especially acknowledge the efforts of the evaluation's technical committee, which was created by the National Planning Commission's then Joint Secretary, Dr. Teertha Dhakal, and includes representatives from the above-mentioned agencies. We thank the World Bank for funding the add-on video assignment intervention. We are grateful for research and logistical assistance provided by Ms. Deepika Shrestha, Mr. Sunil Poudel, Mr. Tri Bikram Pandey, and Ms. Tejkala Uprety of the Center for Policy Research and Consultancy; for research assistance from Ms. Girija Bahety, Ms. Floor de Ruijter, Ms. Kathryn Hirschboeck, Ms. Chhavi Kotwani, Mr. Rayyan Mobarak, Mr. Silver Namunane, and Mr. Jeremy Schlitz; for qualitative research guidance from Dr. Sushan Acharya; and for student assessment development by Ms. Aditi Bhomick, Dr. Alejandro Ganimian, Dr. Peshal Khanal, Mr. Krishna Prasad Adhikari and Mr. Kamal Prasad Acharya.

Summary

In Nepal, as in many low-income countries, student learning outcomes in government primary and secondary schools remain weak. Recognizing that development success requires the country's children and youth to acquire valuable math, science and language skills, the Government of Nepal has prioritized efforts to improve school quality during the seven years of the School Sector Development Program (SSDP), from 2016 through 2023. A key component of the SSDP strategy was a new wave of trainings for 9th and 10th grade math and science teachers, which aimed to improve student learning by: (1) improving teachers' understanding of challenging math and science concepts in the 9th and 10th grade curricula; and (2) encouraging teachers to use new teaching methods involving demonstrations with teaching aids made from locally available materials. Participating teachers attended 10 days of face-to-face training at Education Training Centers (ETCs), after which they were required to complete 5 days of self-study project work, including the creation of 10 lesson plans and related teaching aids.

Rigorous evaluation of teacher training programs is important in Nepal, where the government invests millions of dollars annually in such programs. While evidence from several countries suggests that teacher training programs can improve student learning substantially, evidence from other contexts reveals training program failures, and the evidence base on how to design successful teacher training programs remains thin.

This mixed methods evaluation estimates the impact of the SSDP trainings for secondary math and science teachers on teacher subject knowledge, teaching practices and student test scores and describes the strengths and weaknesses of the programs' design and implementation. It combines a randomized control trial (RCT) of 203 schools in 16 districts with several qualitative research components, including the collection of monitoring data, a "small N" study involving in-person interviews, and a "larger N" part qualitative, part quantitative study involving telephone interviews of teachers and trainers who participated in the SSDP trainings.

We find no evidence that the SSDP trainings for secondary math and science teachers raised student test scores. In fact, our main results allow us to rule out anything more than small positive effects, and in some cases we estimate statistically significant negative impacts. We find weak but suggestive evidence that any negative effects are largest for the students who were highest-performing at baseline. At about $130 per teacher, or $2.60 to $3.00 per student, the cost of the SSDP trainings is similar to that of interventions that have been found to raise student learning significantly in other contexts. We thus conclude that Nepal's policymakers should seek to improve teacher trainings or replace them with more effective interventions.

Drawing on qualitative and quantitative evidence, we describe five sets of problems that may explain why the SSDP trainings did not improve student learning. First, weak governance likely reduced the quality of the ETC training. It appears that trainers were given inadequate time and guidance to prepare training materials, were given no "training of trainers," and in some cases lacked relevant teaching materials. Some ETC's trainers lacked adequate expertise in math and

science. Second, scheduling training sessions on regular school days may have prevented some teachers from participating because substitute teachers were unavailable to teach their classes during the trainings. Teachers' low expectations of the novelty and value of the SSDP trainings may also have lowered participation. Third, we find evidence of serious weaknesses in some teachers' pre-requisite subject knowledge, which may have impeded them from grasping the training content that focused on advanced math and science concepts. Fourth, few teachers seem to have completed the post-ETC self-study project work or adopted new classroom teaching methods, and our evidence suggests two possible explanations: (1) teachers' lack of accountability for the time-consuming development of lesson plans and teaching aids; and (2) teachers' lack of budgets for needed teaching materials. Finally, we find that many students enter grades 9 and 10 with below-grade-level math and science skills. SSDP trainings 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.

Our study has two limitations. First, it is possible that SSDP training impacts grow over time, and we estimated impacts after only one year. Second, the training completion rates in our study schools at endline were unusually low, reducing precision, due to high teacher turnover and low teacher take-up of the training invitations.

Compared to many studies, this evaluation was designed relatively well for external validity, because we use a sample that is nearly nationally representative of all public secondary school students (and the environments in which they live and learn) to study an intervention rolled out through the institutions that were responsible for government training at the start of the study period. A dramatic government reform, however, recently shifted responsibility for basic and secondary education to new local governments, so that trainings identical to those we evaluated will no longer be rolled out, limiting our external validity in a narrow sense. Yet in a broader sense the reform creates valuable opportunity for Nepal's policymakers at all levels to learn from the evidence and pursue improvements in teacher training program design and implementation. The results are also valuable outside of Nepal, as they suggest possible ways of improving the performance of training-center-based in-service teacher training programs.

Given our evidence on the problems that may have reduced the SSDP training program's impacts, we recommend that policymakers experiment with changes of the following sorts: (1) Allocate more training time for methods to identify, and differentiate instruction for, students entering grades 9 and 10 with below-grade-level subject knowledge; (2) Re-design trainings to better accommodate teachers with gaps in pre-requisite subject knowledge; (3) Combine trainings with distribution of related lesson plans and materials (to reduce potential barriers to adoption of new teaching methods); (4) Connect trainings to periodic classroom visits (either in-person or virtual) by mentors or coaches who can advise, monitor and hold teachers accountable for improved teaching; (5) Improve the way trainers are trained, equipped and motivated to deliver high quality trainings; (6) Schedule trainings outside of school hours or during school breaks to increase training program uptake; and (6) Increase efforts to motivate teachers for training by informing them about the novelty and value of the new training.

Table of Contents

Acknowledgements ii
Summary iii
List of tables vi
List of figures vii
List of acronyms viii
1. Introduction 1
2. Intervention 3
2.1 Description 3
2.2 Theory of change 7
2.3 Monitoring plan 7
3. Evaluation 9
3.1 Primary and secondary questions 9
3.2 Design and methods 9
3.3 Ethics 13
3.4 Outcome variables and econometric specifications 13
4. Findings 21
4.1 Intervention implementation fidelity 21
4.2 Impact analysis 23
4.3 Violations of the assumptions underlying the SSDP training program's theory of change 38
4.4 Promising directions for improving program impact 47
5. Cost analysis 52
6. Discussion 54
6.1 Limitations and external validity 54
6.2 Policy and programme relevance: evidence uptake and use 56
6.3 Challenges and lessons 56
7. Conclusions and recommendations 59
Appendix A: Field notes 61
Appendix B: Survey instruments and other evaluation tools (qualitative and quantitative) 64
Appendix C: Pre-analysis plan 65
Appendix D: Sample Design and Calculation of Population Weights 66
Appendix E: Qualitative study reports 69
Appendix F: Supplementary tables 70
References 74

List of tables

Table 1: Description of student attrition ……... 20
Table 2: SSDP math and science training roll-out …… 22
Table 3: Baseline descriptive statistics and balance tests: schools and teachers …… 24
Table 4: Baseline descriptive statistics and balance tests: students …….. 25
Table 5: ITT estimates of impact of SSDP training on students' normalized test scores, full endline and panel samples ……..... 27
Table 6: IV/LATE estimates of impact of SSDP training on students' normalized test scores, full endline and panel samples …… 28
Table 7: Analysis of ITT combined treatment impact heterogeneity by student and household characteristics, full endline sample ……... 30
Table 8: Analysis of ITT combined treatment impact on quantiles of the endline test score distribution …… 32
Table 9: summary of ITT estimates of impact on teacher subject knowledge and attitude …… 33
Table 10: summary of ITT estimates of impacts on teacher attendance …… 34
Table 11: ITT estimates of impacts on math teacher teaching practices (student reports) ……. 35
Table 12: ITT estimates of impacts on science teacher teaching practices (student reports) … 36
Table 13: ITT estimates of impacts on teaching practices: head teacher and teacher reports .. 37
Table 14: Math teacher responses to evaluations of student assessment items …….. 42
Table 15`: Science teacher responses to evaluations of student assessment items…… 43
Table 16: Student performance on below grade-level math questions …….. 48
Table 17: Student performance on below grade level science questions …….. 49

List of figures

Figure 1: Main teacher training intervention theory of change …….. 8
Figure 2: Districts where schools in study are located …… 10
Figure 3: Randomized assignment of the study schools …….. 11
Figure 4: Study timeline …… 14

Acronyms

ETC	Education training center
IRT	Item response theory
ITT	Intention to treat
LATE	Local average treatment effect
MDE	Minimum detectable effect
NCED	National Centre for Educational Development
RCT	Randomized control trial
SEE	Secondary Education Examination
SSDP	School Sector Development Program
SSRP	School Sector Reform Program
TT	Teacher training
VA	Video assignment
VDC	Village development committee
WLS	Weighted least squares

1. Introduction

High quality teaching is critical for successful and inclusive learning (Rivkin, Hanushek and Kain 2005; Chetty, Friedman and Rockoff 2014; Araujo et al. 2016). To improve teaching and learning, nearly all governments invest heavily in teacher training.[1] While some teacher training programs in India, South Africa, and Uganda have induced large learning impacts (Popova *et al.* 2019), others have not increased learning, and even reduced learning for some students, which suggests that designing and implementing successful teacher training programs is difficult (Evans and Popova 2016; Popova, Evans and Arancibia 2016; Loyalka *et al.* 2019). Yet rigorous evidence on the design, implementation and impact of teacher training programs remains scarce.[2] It is, therefore, important to evaluate these programs carefully, so that money is well spent, and policymakers learn how best to design and implement future teacher training programs.

Scrutiny of teacher training programs is especially important in Nepal, where student learning remains weak, despite large expenditures on teacher training over many years. Enrollment rates have risen in recent years, particularly at the secondary level, where the net enrollment rate increased from 35% in 2008 to 66% in 2017 (Ministry of Education 2018). Despite this progress in enrollment, student test scores at both primary and secondary levels remain low. For example, in the 2019 Secondary Education Examination (SEE), held at the end of grade 10, 44.2% of public school students received scores below 2.0 (out of 4.0). Only 4.3% of public school students scored 3.2 or higher, while 40.8% of private school students achieved such scores (Republica 2019). In the National Assessment of Student Achievement of 2018, 32% and 20% of grade 5 students, respectively, performed at below basic level in Mathematics and Nepali (Kafle, Acharya and Acharya 2019). These figures reflect persistent low learning outcomes despite government spending of $21 million on teacher training from 2013 to 2018 (Rauniyar 2019).

This paper reports findings from a mixed methods evaluation of recent trainings for grade 9 and 10 math and science teachers in Nepal's government schools. Teacher training is an important component of Nepal's School Sector Development Program (SSDP), the government's overall plan to raise school quality and inclusivity from 2016 to 2023. Working with key education policymakers in early 2016, we chose to evaluate trainings for teachers of secondary math and science, subjects with very low learning outcomes that are deemed important for attaining Nepal's larger development goals. Relevant government agencies seemed ready to launch these trainings at that time, planning to roll them out gradually across Nepal. It thus seemed feasible to rigorously evaluate this soon-to-be at-scale intervention, with the aim of guiding future policy decisions.

The SSDP trainings seek to improve teachers' understanding of challenging math or science concepts in the 9th and 10th grade student curricula, and to encourage teachers to use methods for teaching these concepts that involve demonstrations with teaching aids made from local

[1] Loyalka and colleagues (2019) report that China spent over $1 billion per year on in-service teacher training, and India spent $1.2 billion on such training between 2012 and 2017. In Mexico, the average teacher spends 23 days per year in teacher training, and between 2000 and 2010 "nearly two thirds" of World Bank supported education projects included teacher training (Popova, Arancibia and Evans 2016).
[2] Popova, Evans and Arancibia (2016) found only 23 papers on teacher training in developing countries.

materials. They required teachers to attend 10 days of face-to-face training at education training centers (ETCs) and, after returning to their schools, to complete five days of self-study project work, including creation of 10 lesson plans and related teaching aids.

Though focus primarily on the SSDP teacher training, we also developed an add-on video assignment for teachers in half the schools assigned to the teacher training. This assignment required training participants to submit videos of themselves implementing in their classrooms lesson plans that they were required to make as part of the SSDP self-study project work. In the end, the study lacked power to distinguish impacts between the two variants of the SSDP training (with and without video assignment), and since adding the video assignment was a very small change in the intervention's design, this evaluation of the SSDP training's impacts reports average effects over these two variants of the training. Online Appendix B discusses the video assignment further.

This study combines a randomized control trial (RCT), which involves a sample of 203 schools in 16 districts, with qualitative methods designed to complement the quantitative study. The sample is nearly nationally representative of all public secondary school students and the environments in which they learn. We implemented a preliminary qualitative study in 16 schools in 4 districts in early 2017 to deepen our understanding of context, refine our evaluation questions, and explore ways to measure key variables. We also collected monitoring data during the intervention's roll-out through frequent phone contact between the research team and ETC personnel (Shrestha 2019). To examine the nature and quality of the roll-out, we added two more qualitative studies: (1) a "small N" study using in-depth in-person interviews of teachers, trainers and other actors in three study districts (Acharya and Upreti 2019); and (2) a "larger N" part qualitative, part quantitative study based on telephone interviews with 98 teachers across all study districts who had attended the SSDP trainings, and with 23 trainers (Schaffner, Glewwe and Sharma 2019a).

The main evaluation questions that guided the study's design are:

- What are the impacts on student learning outcomes of the SSDP teacher trainings for 9th and 10th grade math and science teachers?
- What are the impacts on teacher subject knowledge and teaching practices of these SSDP teacher trainings?
- What assumptions underlie the SSDP teacher training program's theory of change, in what ways might these assumptions be flawed, and what are the resulting strengths and weaknesses of the design and implementation of the SSDP teacher trainings?

Suspecting that trained teachers are more motivated to implement new teaching practices in schools with better school management (i.e. where head teachers or other actors hold teachers more accountable and provide them better support), we sought to answer a fourth question:

- How do SSDP teacher training impacts differ across schools with stronger and weaker initial school management?

A review of the education and economics literatures since 2000 reveals only 10 studies of the impact of teacher training programs on student learning in developing countries that use credible methodologies.[3] Three of these are from Kenya (Lucas *et al.* 2014; Jukes *et al.* 2017; Piper *et al.*

[3] For details on the search methods, see Damon and colleagues (2019). We include only studies involving teacher training, and that were published or in high-profile working paper series from 2000 to 2018. We consider three methodologies as credible: RCTs, difference in differences, and regression discontinuity.

2018),[4] two are from China (Loyalka *et al.* 2019; Lu *et al.* 2019), and one each are from Argentina (Albornoz *et al.* 2019), Mongolia (Fuje and Tandon 2018), Papua New Guinea (Macdonald and Vu 2018), the Philippines (Abeberse, Kumler and Linden 2014), and Tonga (Macdonald *et al.* 2018). The teacher training programs examined are diverse. For example, the programs in the Philippines and Mongolia involved only two and three days of training, respectively, while others had 12 days (Lucas *et al.* study of Kenya and Uganda) or 15 days (Loyalka *et al.* China study). Several programs had follow-up coaching and/or workshops, while others did not. All but two were for primary school teachers; the only exceptions are the Albornoz and colleagues Argentina study and the Loyalka and colleagues China study. Finally, about half of the programs combined teacher training with new curriculum and/or pedagogical materials, while others (the two China studies, and the Argentina, Papua New Guinea and Tonga studies) had only teacher training.

The studies suggest substantial diversity in program impacts. Most found small positive impacts on student learning of 0.1-0.2 standard deviations of the distribution of student test scores. Yet others found larger effects, up to about 0.6 standard deviations in the Jukes and colleagues (2017) Kenya study and the Argentina study. Still others found no impact (both China studies).

This study adds to this small literature in three ways. First, we study training at the secondary level, for which there is currently only one study (China study of Loyalka *et al.* 2019). Second, we examine a program at national scale, using a sample that is close to nationally representative. Only two of the previous studies were on a national scale: the Mongolia and Tonga studies. Finally, we evaluate a government program, rather than a program designed by a research team or non-governmental organization. Of the ten previous studies, only the Mongolia study, the two China studies, and the Piper *et al.* (2018) Kenya study evaluated government programs.

Section 2 describes the intervention's main features and its theory of change. Section 3 describes the evaluation design. Section 4 presents findings of the intervention's impacts, and the strengths and weaknesses of its design and implementation. Section 5 details intervention costs. Section 6 presents limitations and external validity. Section 7 concludes and provides recommendations.

2. Intervention

2.1 Description

The teacher training (TT) intervention invited all 9th and 10th grade math and science teachers in selected government schools to enrol in relevant SSDP trainings. An initial SSDP document (Ministry of Education 2016) motivated the need for new and improved teacher trainings by noting widespread beliefs that previous teacher trainings had not transferred effective teaching methods from the trainings to the classrooms. It especially highlighted the apparent failure of the "needs-based approach" of the previous wave of trainings, which encouraged individual Educational Training Centers (ETCs), of which there were then 29 across Nepal, to develop training curricula tailored to local teachers' requests. The SSDP trainings were, therefore, to be developed in a more centralized fashion by the National Centre for Educational Development (NCED).

[4] The Lucas and colleagues (2014) paper also includes analysis of data from Uganda.

The official curriculum for the trainings covered challenging math or science concepts in the 9th and 10th grade curricula, and specific demonstration-based methods (often using teaching aids made from local materials) to teach specific concepts. Participants attended 10-day sessions at ETCs, and then were expected to complete five days of self-study project work that included completion of: a) 10 lesson plans; b) an action research project related to a classroom or school problem; and c) two of several specified activities.[5] 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 otherwise were offered no monetary incentives for attending. 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. Adequate scores were required for teachers to obtain credit for the training in their general performance review records.

The TT intervention that rolled out in the study schools differed in small ways from the training that was 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 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 previous education plan (the School Sector Reform Program or 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.[6] Third, while the SSDP trainings were designed to include two modules, each with 10 days of ETC training and five days of project work, in practice only the first module has been rolled out. The training roll-out throughout the country has stalled, because of a dramatic "federalizing" reform of government institutions that accelerated in September of 2018, when the ETCs were shifted from federal-level government administration to administration by the new provincial governments.

The curricula for the SSDP math and science trainings were described in two documents sent by the NCED to the ETCs (NCED 2017a and 2017b). These documents are brief (11 pages each), and the NCED faced challenges in exercising oversight of the trainings at local ETCs, thus we conducted a telephone survey of 98 teachers (from all study districts) who had attended SSDP

[5] 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 rates; or (4) use a clinometer to calculate height and distance for objects near the school. For science training, the 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 use for teaching about constellations.

[6] While this means that the study intervention differs somewhat from the SSDP intervention in non-study districts, we believe that the study is very relevant for policy discussions in Nepal. Indeed, there is precedent for providing government training to non-permanent teachers. Moreover, 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 of non-permanent teachers.

trainings and 23 trainers (from 12 of the 14 ETCs serving our study districts) to learn the *de facto* content and training methods used during the trainings (Schaffner, Glewwe and Sharma 2019a).

The descriptions of training content from the telephone interviews are broadly consistent with the official documents, though details vary across the ETCs. The interviews confirm that the sessions devoted significant time to helping teachers learn 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 them 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 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). Among 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 (where responses were more varied). In answer to other open-ended questions, teachers mentioned making and using litmus paper and using special seating methods to teach about sets or the periodic table of elements. Some, but not all, trainings appeared also to emphasize the pedagogical practice of having students work in groups; over one third of teachers and 9 out of 23 trainers mentioned group work or pair work as a discussion topic. For details, see Schaffner, Glewwe and Sharma (2019a) and Section 4.1 below.[7]

Target populations. The intervention's immediate targets were teachers of 9th and 10th grade math and science. Our baseline data show that these teachers are almost all male (98% for math and 89% for science) with average age of 35 years.[8] Only 45% of math teachers and 24% of science teachers had previously lived in their current schools' communities. Approximately 81% have at least Bachelors' degrees in math or science, and 46% of math teachers and 36% of science teachers have at least Masters' degrees (in any field, including education). About one fourth of the teachers (30% for math and 23% for science) have permanent positions that are funded by the federal government; the rest have diverse positions that are less secure and differ in funding source and salary. Among math (science) teachers, 31% (24%) report receiving math (science) teacher training under the SSRP (the previous education policy plan) and 70% (59%) report receiving any past in-service math or science teacher training from government or non-governmental sources. During the study period, because of uncertainties related to the federalizing reforms of Nepal's government institutions, little other government training was available for 9th and 10th grade math and science teachers. Some non-governmental organizations (NGOs) may have offered trainings during this period, but the reach of these

[7] The content of the teacher trainings actually rolled out differed from the content we expected based on the SSDP document we had when we planned our evaluation. The SSDP document (Ministry of Education 2016) and conversations with policymakers at that time suggested that the trainings would focus on pedagogy for child-centered learning, inclusive education, formative assessment and differentiated teaching to meet student needs (thus, our baseline instruments give special attention to these topics). In practice, there were delays in defining the content for the trainings, and their focus was shifted to the demonstration-based teaching methods described above. Our endline instruments were designed for the revised training content.
[8] For statistics reported in this section, we use baseline data and population weights. The results are roughly representative of Nepal's government schools with at least grades 1 to 10 (see the methods section below).

organizations is limited, and their focus tends to be on primary education. At endline, only 14% of math teachers reported having ever had an NGO math teacher training, and only 9% of science teachers reported having ever had an NGO science teacher training.

The schools in our sample include at least basic (1 to 8) and secondary (9 and 10) grades. Such schools educate 97% of Nepal's 9th and 10th grade public school students (based on education management information system statistics). Our baseline data show that these schools are large, with 86% having over 200 students. Yet the typical school is also quite remote, with only 22% within a 15-minute walk of a motorable road and 38% more than a 3-hour walk from a motorable road. Median class sizes (number of students per "section") were 43 in grade 9 and 39 in grade 10 at baseline, but vary greatly, with sizes at the 95th percentile of 97 (79) for grade 9 (grade 10).

The intervention's ultimate targets are 9th and 10th grade students, who were in grades 8 and 9 at baseline. More than half are female at baseline, 55% in both 8th and 9th grade, perhaps reflecting that boys are more often sent to private schools. Data for 9th graders at baseline indicate low levels of parental education; students reported that only 83% of fathers and 60% of mothers are able to read and write, and only 28% of fathers and 12% of mothers obtained at least some secondary education (grade 9 or higher). Asset data suggest that while typical secondary students' families are not among Nepal's poorest, they are poor by global standards. Most families of 9th graders have mobile phones (95%), 55% have televisions and 36% have bicycles, but only 13% have refrigerators and 10% have computers. Our endline data show that 40% of 10th graders are Brahmins or Chhetris, 36% are Janajatis, 13% are Dalits, 7% are from Terai or Madhesi castes, 3% are Newaris, and less than 1% are Muslims.[9] Importantly, head teacher and teacher responses to qualitative questions suggest that many students enter grades 9 and 10 with skills significantly below grade level. We examine this challenge more closely below.

Impact heterogeneity. As indicated in our pre-analysis plan (Appendix C), SSDP training impacts may be greater for teachers who did not participate in the previous (SSRP) training and for teachers with non-permanent appointments (who can be dismissed for inadequate performance). Impacts may also differ by teachers' years of experience, although we have no strong prior regarding the role of teacher experience; teachers with less experience may gain more from the trainings, since they have had less time to learn on the job, but they may also gain less from the trainings, since their pre-service education may have been superior to that of earlier cohorts. We were especially interested in whether training impacts are larger in schools with higher quality school management. The school management quality indices we use are described in the appendix of our pre-analysis plan (Appendix C). Our baseline data suggest that most manage-ment tasks are done by head teachers, who may visit classrooms, give feedback, convene meetings and provide other practical leadership. Head teachers' management inputs seem to vary widely across schools, in part because some head teachers have more management training and lower teaching loads than others. We hypothesize that teacher trainings have larger impacts on teaching practices and students' learning when high quality management does more to support new teaching methods and hold teachers accountable for using new

[9] For baseline data we must infer ethnicity/caste/religion from surnames, while at endline the head teacher reports ethnicity for each student. We use endline data here, because we believe them to be more accurate.

teaching methods. Program impacts on learning may also differ by student types. Of special concern is that the trainings, which focus on improving the teaching of advanced grade 9 and 10 concepts, may have no effect on students who enter those grades with below-grade-level subject knowledge. We also check for impact heterogeneity by student gender, ethnicity, and parental education and assets.

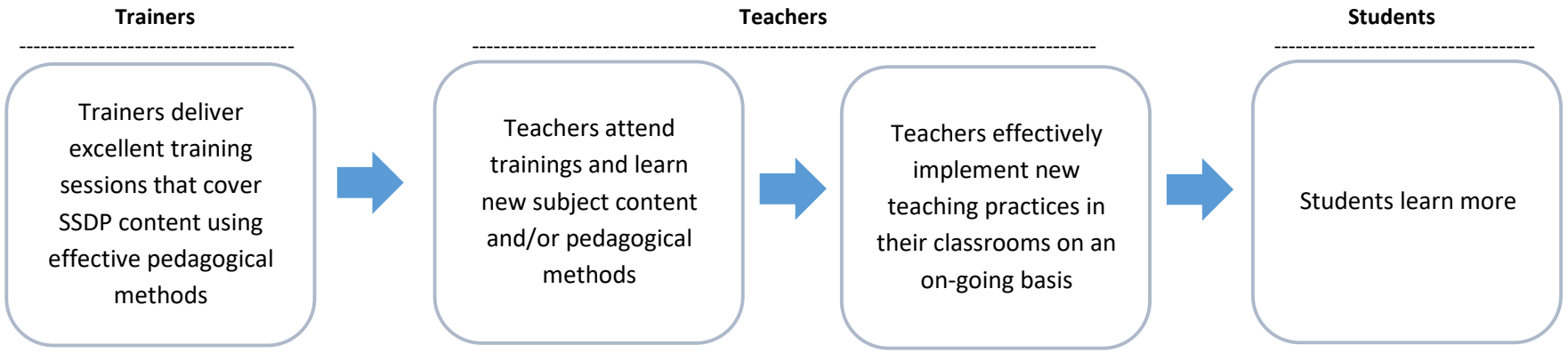
2.2 Theory of change

The approach we took to articulating the theory of change and identifying its underlying assumptions was to (a) identify all the decision-making actors along the logical chain linking the SSDP training policy to its ultimate objective of improving student learning, and for each of these actors asking (b) how must they respond to the new opportunities created by the policy (or what roles must they play) for the policy to improve student learning, and (c) what might they lack in motivation, resources or capacity, preventing them from responding in ideal ways? We focused on trainers (who must conduct high quality trainings), teachers (who must attend and learn from the trainings, and then implement improved teaching practices in their classrooms on a daily basis), and students (who must study and have the potential to learn better from the improved teaching practices). We illustrate the theory of change, and list the critical assumptions underlying the theory of change, in Figure 1. We examine these assumptions in Parts 4.1, 4.2.4 and 4.3. Part 4.1 uses administrative data to examine the training session roll-out and teacher attendance at the trainings. Part 4.2.4 estimates program impacts on intermediate outcomes related to teacher subject knowledge and teaching practices, as well as final student learning outcomes, while Part 4.3 uses mixed methods to examine the motivation, resources and capacity that trainers, teachers and students bring to their roles in generating training program impact.

2.3 Monitoring plan

Because of weak monitoring practices within government teacher training institutions, we devised our own monitoring activities. We collected data on training session dates, numbers of trainers involved, teacher invitation activities, and teacher attendance through frequent phone calls by research team members to the ETCs (see Shrestha 2019). Unfortunately, we could not obtain any scores that the ETCs might have recorded on the quality of teachers' participation during the trainings. These calls also provided qualitative information on the difficulties ETCs experienced in rolling out the trainings (and video assignment recording sessions). In addition, we designed two qualitative studies, and some endline survey questions, to understand the nature and quality of the training sessions and of teachers' completing their self-study project work. While the research team's involvement in monitoring may have improved governance of the program, thereby reducing the external validity of the study, we believe this effect was very small, because the ETC personnel we spoke to did not seem to consider themselves accountable to us. In fact, repeated phone calls were needed to obtain even basic data, and one ETC provided no information.

Figure 1. 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

3. Evaluation

3.1 Primary and secondary questions

As indicated in the introduction, the main evaluation questions that guided the study's design are:

- What are the impacts on student learning outcomes of the SSDP teacher trainings for 9th and 10th grade math and science teachers?
- What are the impacts on teacher subject knowledge and teaching practices of these SSDP teacher trainings?
- What assumptions underlie the SSDP teacher training program's theory of change, in what ways might these assumptions be flawed, and what are the resulting strengths and weaknesses of the design and implementation of the SSDP teacher training?

Our secondary question is:

- How do the SSDP teacher training impacts differ across schools with stronger and weaker initial school management?

3.2 Design and methods

We designed an RCT to estimate SSDP training impacts on intermediate and final outcomes, and designed several qualitative studies to complement the RCT.

RCT sample design. Our RCT sample design was shaped by four objectives, which we identified through conversations with our government collaborators. First, the sample should be large enough to yield sufficiently precise impact estimates. Conservative power calculations suggested the need to include about 100 treatment and 100 control schools to estimate the impact of the TT treatment on student test scores with adequate power.[10,11] Second, the sample should be approximately representative (using population weights) of all schools in Nepal that have at least grades 1 through 10. This would make the data useful for describing national secondary education challenges about which little was known, and policymakers also deemed it important to include districts from all major regions of the country. Third, because, during the development of the SSDP teacher training curriculum, it seemed that there could be substantial overlap between the content of SSDP trainings and that of SSRP trainings, we stratified the sample to over-sample schools

[10] We aimed for a sample of schools 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). The estimate of the standard deviation of a test score variable was about 20 and the estimate of the Intra-Cluster Correlation Coefficient was around 0.65, which is very high and implies the need for a large sample of schools. We anticipated being able to obtain more precision by using baseline test scores as controls in endline impact regressions.
[11] Budget constraints precluded inclusion of an additional 100 schools to estimate the impact of the training intervention with the video assignment. Thus we chose to randomly allocate half the 100 treatment schools to also receive the video assignment add-on intervention. While this approach did not guarantee adequate power to distinguish the effects of the TT intervention with and without the VA, we judged that the potential for learning from qualitative and quantitative study of the VA justified adding the low-cost VA to the study.

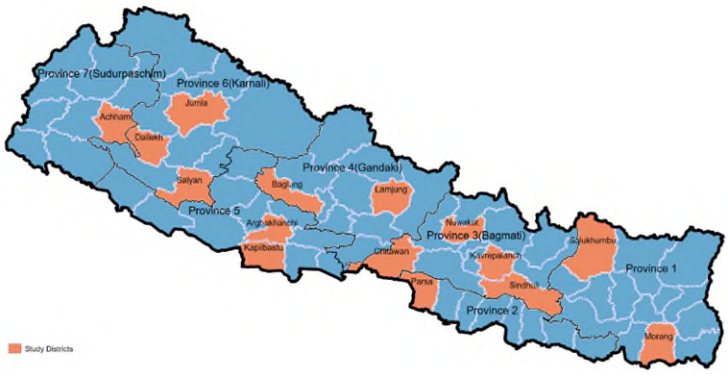
where few teachers of grade 9 and 10 math and science had completed SSRP training. Fourth, the sampling procedure should minimize the potential for spillover effects to lead to biased impact estimates. The following paragraphs describe the sample selection process in more detail.

To obtain a sample representative of most of Nepal, while containing costs, we chose a two-stage design, first sampling districts and then sampling schools within districts. To reduce data collection costs, we (in consultation with our Nepalese government partners) eliminated from consideration 10 of the most remote or otherwise difficult districts. From the remaining 65 districts (where 94.3% of Nepal's schools with at least grades 1 through 10 are located), we randomly selected 16 districts, and then sampled schools only within those 16 districts. Schaffner, Glewwe and Sharma (2018) describe the sampling in detail. Figure 2 shows the 16 selected districts.

We sorted schools within districts into two strata, "priority" and "non-priority," and over-sampled the former. Priority schools were defined as those with no evidence (from NCED hard copy records) of any teacher with permanent or unknown contract type who completed all three SSRP training modules. This rule was dictated by the idiosyncrasies of the available records. Further details are given in Schaffner, Glewwe and Sharma (2018). Within each district, we selected two-thirds of our sample schools from the priority stratum, and one-third from the non-priority stratum.

To facilitate selecting two-thirds of the sample within a district from the priority stratum, it was useful to choose a number of schools per district that is divisible by three. To allocate one fourth of the sample each to the TT treatment without Video Assignment (VA) and the TT treatment with VA (while allocating the other half to the control group), it was useful to choose a number divisible by four. Therefore, we selected 12 schools per district in most study districts. At the request of government partners, we doubled the number of schools in Morang, one of the larger districts.[12] Thus the aim was to select a total sample of (12×15 + 24 =) 204 schools. In the end, it included only 203 schools, since Solukhumbu district had only 3 non-priority schools.

Figure 2. Districts where schools in the study are located

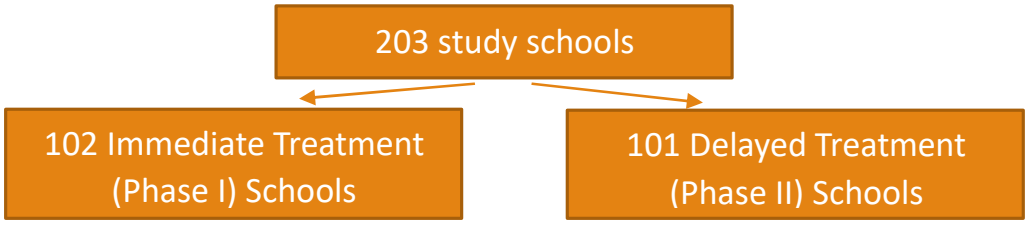


Note: This map was generated using open access files at https://opennepal.wordpress.com/ and https://gadm.org/data.html.

[12] Morang was given a 'double' sample because it is the largest of the 16 districts; in the administrative data used to select schools, Morang has 154 of the 1,334 schools in the 16 districts, more than any other district.

Concerned about potential spillovers of impact from treated to untreated schools, we selected priority and non-priority schools within districts in a way that reduced the probability of any two sample schools being near each other. We first grouped schools (within districts) by the Village Development Committee (VDC) areas to which they belonged. The average VDC had 1.7 eligible schools.[13] We then sampled VDCs, and randomly selected only one school per VDC.

Figure 3: Randomized assignment of the study schools



Assignment of schools to study arms. As shown in Figure 3, the primary randomization divided the 203 study schools, within district and priority stratum, into two groups of equal size: (1) "Phase I" schools, which were to receive the SSDP teacher training in late 2017; and (2) "Phase II" schools, which were to receive that training only after May of 2019, and which serve as the control group during the study period. To minimize the potential for spillover effects of training onto Phase II study schools, training was also to be delayed until after May of 2019 for non-study schools in the same VDCs as the Phase II schools. Random selection was done without replacement, using a random number generator in STATA 13. Within schools, we tried to include all teachers of grade 9 and 10 math or science, and all grade 9 and 10 students.

Randomly allocated assent and test administration procedures. Within each district-study arm combination, we randomly assigned one-third of schools to a student informed assent process that was implemented before the assessments (as is standard, and as done at baseline, but which may have made the low stakes nature of the assessments more salient to students), while allocating the other two-thirds of schools to an assent process administered immediately after the assessments (and before students submitted their assessment papers, giving them the opportunity to choose not to submit). Following Institutional Review Board directives, whether conducted before or after the assessments, the assent process made clear that the assessments would not count toward the students' grades in school and that their scores would not be revealed to anyone at their school. We chose to delay the assent process until after students took the assessments after observing low scores and enumerator reports of poor assessment-taking discipline at baseline. We kept the baseline assent process in one-third of the schools, however, in order to evaluate whether changing the assent process improved test performance. According to data gathered during an enumerator debriefing, all students assented, although 5 students "ran away" between assessments. Procedures for asking teachers or head teachers to encourage students to do their best on the assessments were also strengthened at endline in all schools.

[13] Despite the intention of government collaborators to delay the roll-out of the SSDP trainings in the 16 selected districts until after baseline data were collected, a few such trainings did occur in these districts.

In addition, within each district-study arm combination, students in half the schools were assigned to take the math assessment first, while the other half took the science assessment first. At baseline all students took the math assessment first. Since test-taking fatigue may reduce performance on the second test vis-à-vis the first, and so reduce the quality of the second test, this randomization equalizes assessment quality across the two tests.[14]

Population weights. To produce estimates of the mean or variance of a population characteristic, or the average of a heterogeneous effect, for the population of schools (with at least grades 1 to 10) 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 calculate these weights using Monte Carlo methods.[15] When using these weights, we interpret our sample as "nearly" nationally representative. It is not fully nationally representative because: (1) the sample frame excludes 10 remote districts; and (2) our weights do not adjust for small departures from using sampling probabilities proportional to size when selecting the 16 districts.

Data collection instruments for RCT study. We gathered endline data using head teacher, teacher, School Management Committee and student questionnaires; student assessments in grade 9 and 10 math and science; teacher evaluations of student assessment items (which allow indirect assessment of teacher subject knowledge); student tracking forms; teacher turnover forms; assessment administration conditions forms; and enumerator debrief forms. Some regressions also use baseline student assessment scores; school-, teacher-, or student-level covariates from baseline questionnaires; or monitoring data on teacher participation in the SSDP trainings. All questionnaires and forms are available in Appendix B.

Monitoring data collection from ETCs. Research assistants made frequent phone calls to the ETCs during program implementation to: get updates on training dates; obtain reasons for delays; ensure compliance with protocols for inviting teachers from treatment schools and for not inviting teachers from control schools and neighboring schools; and gather data on training curriculum and attendance. Our econometric analysis relies on these data, rather than teachers' self reports, regarding their participation in the SSDP training. We found the self reports to be inaccurate, probably because the names that teachers recognized for specific trainings were different from the official names, and many teachers attended multiple trainings over the years. For more on monitoring of training rollout see Shrestha (2019), which may be found in Appendix E.

Telephone interview study. To obtain more detail on the nature and quality of de facto SSDP training than would be collected in the endline quantitative survey, and to examine more elements of the theory of change, we sought to interview by phone all teachers from the quantitative sample who were in treatment schools at baseline and who completed the SSDP trainings. Of the 221 teachers of grades 9 and 10 math or science in the schools assigned to the treatment group, we have baseline data for 192. Of these, 120 completed SSDP training, and we were able to interview

[14] We detected no impact of the timing of the assent process, nor of the order of the tests, so we do not report these results in any detail in this report.

[15] Monte Carlo methods were needed to account for the complex structure of sampling without replacement. The details of how these weights were constructed are shown in Appendix D.

98 by telephone. We also sought to interview one math trainer and one science trainer in each of the 14 ETCs serving our 16 districts. We ultimately interviewed 12 math trainers and 11 science trainers. The interview protocols (see Appendix B) had both closed ended and (short answer) open ended questions. The final report (Schaffner, Glewwe and Sharma 2019a) is in Appendix E.

In-depth in-person interview study. To confirm and deepen our understanding of the telephone interview study results, we commissioned a small-N qualitative study (Acharya and Upreti 2019). After consulting our government partners, they selected three of our quantitative study districts, one each in Nepal's Eastern Terai, Western Mid Hill and Far Western regions. In each district, interviews were conducted with two math and two science teachers from the quantitative sample who were interviewed at baseline in treatment schools and had completed the SSDP trainings, plus another teacher (a local teachers' federation representative). In each district the researchers also interviewed one trainer involved with the SSDP math training and one involved with the SSDP science training. Interviews lasted about 60 to 90 minutes per interviewee. The two researchers were in frequent contact while in the field, to maintain consistency. They recorded the interviews with permission and transcribed their notes after returning to Kathmandu. They jointly developed a system to color-code the data to organize and systematize their impressions. The interview protocols are in Appendix B and the final report (Acharya and Upreti 2019) is in Appendix E.

Strategies to avoid bias. We sought to avoid spillover effects by sampling schools in a way that reduced the likelihood of a control school being near a treatment school. We attempted to avoid bias due to Hawthorne and John Henry effects by presenting the study to all participants as part of an effort to "improve learning in secondary schools" rather than as an evaluation of the teacher training program, and by including questions about teacher training only at the end of questionnaires. While all study participants were aware that they were part of a study, our procedures aimed to prevent respondents from thinking of themselves as members of treatment or control groups, thereby preventing them from responding to observation differently.

Evaluation timeline. Figure 4 shows the evaluation timeline, which includes rows for intervention rollout phases, important contextual factors, and qualitative and quantitative research stages.

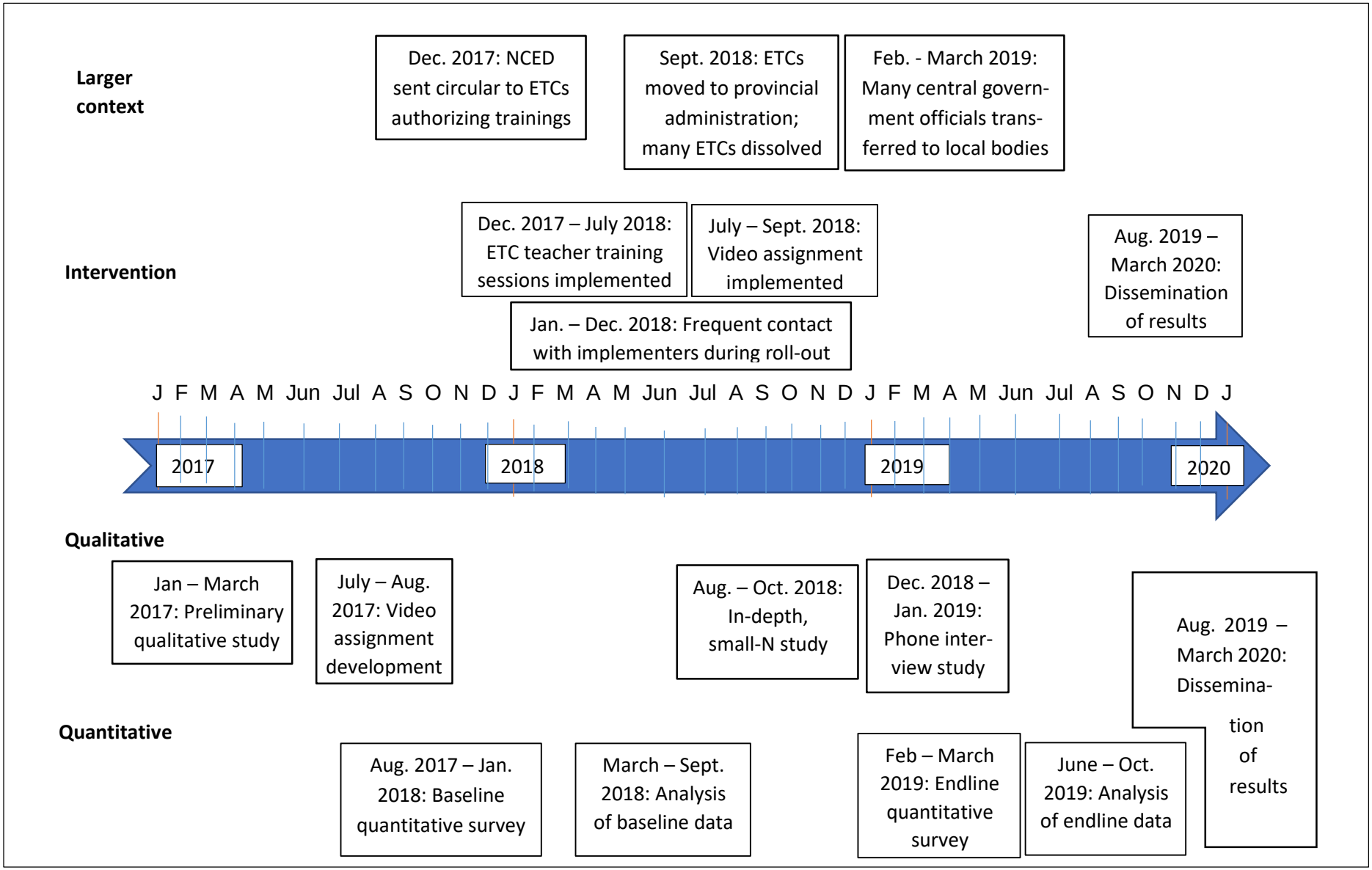
3.3 Ethics

All research protocols and instruments were reviewed and approved by Institutional Review Boards at Tufts University and the University of Minnesota. The relevant protocols are 1707008, 1807017, and 1901005 at Tufts University, and STUDY00000915 at the University of Minnesota.

3.4 Outcome variables and econometric specifications

The goal of the RCT is to estimate intention to treat (ITT) and local average treatment effect (LATE) impacts of the treatment on the final outcome, student learning, and on the intermediate outcomes of teacher subject knowledge and teaching practices. With only a few exceptions (all noted in the text or footnotes), we use measures and methods reported in our pre-analysis plan.

Figure 4. Evaluation Timeline



Final outcome variables. Almost all grade 9 and 10 students present on the day of endline assessment completed two one-hour assessments, one in math and one in science.[16] U.S.-based consultants with psychometrics expertise and familiarity with international assessment item banks drafted the assessments. We provided them English translations of student textbooks and SSDP teacher training curriculum documents, and asked them to construct assessments tailored to the Nepalese curricula for these grades and subjects, giving special attention to curriculum content emphasized in the SSDP trainings, while including items at lower grade levels (to assess how many students enter grades 9 and 10 with skills below grade level). The consultants, as asked, included some items from the baseline assessments (to link with those assessments), but drew most of the questions from international assessment item banks, allowing incorporation of high-quality items that had been refined by intensive pre-testing. Two assessment versions (Versions "A" and "B") were drafted for each subject and grade, to reduce the risk of students copying from each other (to give students sitting in rows in crowded classrooms alternating exam versions) and to increase the subject content covered by the assessments. The two versions included common items to link between them. The drafts were reviewed, amended and translated into Nepali by local assessment consultants, to ensure the tests' relevance to Nepal's curriculum and testing style.[17] After pre-testing, 6 items with the lowest correct response rates were dropped from each assessment to produce the final assessments with 35 items each. All items are multiple choice. Unfortunately, printer errors caused some assessment copies to have the wrong page for the "A" versions of the grade 9 and 10 math assessments, which were distributed in a few schools before the problem was detected. We treat these as a third version ("Version C") of the relevant math assessments. Fortunately, most of their items link them to the correctly printed assessments.

For all students we estimate, separately by grade and subject, indices using item response theory (IRT) methods to measure overall math achievement and overall science achievement. For each grade and subject, we linked the two or three assessment versions using a 2-parameter logistic IRT model.[18] Using the assessment consultants' item maps, we then identified the items closest to the content emphasized in the SSDP training. We used the same IRT methods to produce grade- and subject-specific achievement indices in these more specific content areas.

Intermediate outcomes. We examine impacts on a range of intermediate outcome variables, measuring teachers' subject knowledge, attendance, and teaching practices. At endline, to assess teachers' subject knowledge without explicitly asking them to take assessments, we asked teachers of grade 9 and 10 math and science to fill out anonymous evaluations of selected student assessment items. The evaluation forms showed 12 items from the student assessments to the teachers, and then asked them to: (1) rate their clarity; (2) provide the answers they thought the items' writers intended as the correct answers; (3) estimate what fractions of their students would answer the items correctly; and (4) rate how well-tailored the items are to the curriculum used in

[16] Appendix A gives the protocol that determined which students were included in the endline assessments.

[17] The local consultants had also prepared the baseline assessments.

[18] Our pre-analysis plan stated that we would estimate 2- and 3-parameter logistic IRT models, using a 3-parameter model only if estimation was feasible and a likelihood ratio test rejected the 2-parameter model. In practice, we could not estimate a 3-parameter model as the algorithm did not converge. Online Appendix A, Tables A4 through A7, show the discrimination and difficulty parameters for the 2-parameter model.

Nepal.[19] We created teacher subject knowledge indices, using IRT methods, using teachers' answers for the response options they believed the item designers intended as the correct ones.

We measured teacher attendance three ways, each having strengths and weaknesses. First, we instructed enumerators to interview all grade 9 and 10 math and science teachers on the first day they visited a school, and to record which teachers were absent. (In some cases, enumerators interviewed absent teachers on the second day of a school visit.) We believe that this accurately describes teacher presence or absence the first day a school is visited but it may overstate typical teacher attendance, because all school visits were arranged in advance. Second and third, we asked head teachers and students about teachers' attendance rates. Head teacher ratings were obtained only for teachers who were not the head teacher. We do not expect bias in student responses for teacher attendance, but the use of broad attendance rate categories and the need for students to report for a long recall period may reduce the precision of the student measure.[20]

Our teaching practices measures are derived from endline head teacher, teacher, and student responses to questions on teachers' practices, which we describe when we present the impact estimates in Section 4.2.4. Many teaching practice outcomes are ordinal, with categories for Likert-scale opinions or activity frequencies; some are dichotomous, while others have 3 to 7 categorical options. We use probit models to estimate impacts on dichotomous outcomes. To estimate impacts on 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, and then use ordered probit models. For brevity, we describe our estimation methods only for continuous outcome measures, such as test scores, for which we use linear regressions; these methods can be easily adapted for use with dichotomous or ordinal outcomes.

Student-level ITT impact regressions. ITT impact estimates assess the impact on final or intermediate outcomes of inviting all teachers of 9th or 10th grade math or science in a school to the SSDP trainings. The main regression equation for student-level outcomes, which we estimate for all endline sample students (including those without baseline data), has the form:

$$Y_{is1} = \beta_0 + \beta_T Treat_s + A_s\beta_A + S_s\beta_S + \varepsilon_{is1}$$

[19] The 12 items were selected by the U.S.-based consultants who drafted the endline student assessments. They were asked to select two items from lower grade levels (so that we could start the assessment with items that would not look daunting) and then pick 10 9th and 10th grade items that correspond to topics covered in the official SSDP curriculum outline. We also asked them not to select items with large graphical elements, because we wanted to keep the assessment document short to encourage participation.

[20] Our pre-analysis plan also called for measuring teacher attendance through inspection of school attendance registers. We gathered these data but chose not to use them, because two patterns in the data suggested that they contained significant errors. First, attendance rates as recorded in the registers rise from 77% on the day of the visit to 88% five days before the day of the visit, consistent with the hypothesis that, once they return after an absence, teachers may adjust the logs to indicate themselves present on preceding days. Second, while the attendance logs showed only 77% of teachers to be present on the day of the school visit, the interviewers' efforts to track down teachers on the first day of the school visit showed that in fact 90% of them were present, suggesting that teachers who are present do not rigorously record their attendance each day.

where Y is a student learning measure, Treat is a dummy variable for schools randomly selected for the trainings, A is a vector of indicators showing 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.[21]

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:

$$Y_{is1} = \beta_0 + \beta_1 Y_{is0} + \beta_T Treat_s + A_s\beta_A + S_s\beta_S + \varepsilon_{is1}$$

using the same notation as above, and the subscript 0 refers to baseline. The main specifications are estimated by weighted least squares (WLS), using the weights described above.

School-level ITT impact regressions. The main regression equations for school-level outcomes, which use endline data only, have the form

$$Y_{s1} = \beta_0 + \beta_T Treat_s + S_s\beta_S + \varepsilon_{s1}$$

where Y is a school-level outcome variable, and the rest of the notation is as above (now without student subscripts). The main specification is estimated using weighted least squares.

Teacher-level ITT impact regressions. The main regression equations for studying teacher-level outcomes, which use WLS and only the endline sample, have the form:

$$Y_{ts1} = \beta_0 + \beta_T Treat_s + S_s\beta_S + \varepsilon_{ts1}$$

where Y is a teacher-level outcome, t is the teacher subscript, and all other notation is as above.[22]

LATE estimates. ITT impacts may be low in part simply because invited teachers do not take the training or because trained teachers leave treatment schools and possibly move to control schools soon after their training. ITT impacts are relevant for assessing the impacts of policy interventions that cannot force participation, yet we also estimate LATE impacts, which measure the average effects of teachers' actual *receipt* of treatment.[23] We matched students to their math and science

[21] Following our pre-analysis plan, we estimated variants of all equations described here that allowed the intervention's impact to differ between treated schools where treated teachers were and were not required to do the video assignment. After adjusting for multiple hypothesis testing, no differences in impact were significant. We present some of these results in Online Appendix A.

[22] Our pre-analysis plan stated: "If teacher interview data are missing for more than 5% of a school's 9th and 10th grade math and science teachers, we will adjust the weights to account for uneven non-response across schools, multiplying it by the ratio of the total number of relevant teachers in the school to the number of relevant teachers for which interview data are available." We no longer believe this is a well-motivated adjustment and did not try it. The purpose of the adjustment was to make the estimates more representative of the full population of teachers. The teachers for which responses are missing, however, are often the head teachers, who are likely to be systematically different from other teachers.

[23] LATE estimates measure the average impact on students of having a teacher who was trained, averaged over the students whose teachers were trained; these estimates may not measure the impact that would have occurred for students of teachers (in treated schools) who did not participate if they had participated.

teachers, using student and teacher reports of the section names to which they belonged.[24] We instrumented the indicators of teachers' participation in the training by their schools' treatment assignment, both for the full endline sample, not controlling for baseline scores, and (for student test scores) for the panel sample, controlling for baseline scores. The LATE regression equations for student- and teacher-level outcomes are as above, except "Treat" (indicator of assignment to treatment) is replaced by an indicator of training participation and is instrumented by "Treat".

Estimating heterogenous impacts. We also examine impact heterogeneity by adding (one at a time, in our ITT regressions) interaction terms between the treatment indicator and measures of the dimensions of heterogeneity discussed above. (We also include un-interacted heterogeneity variables.) We do this for the following variables: (1) extent of previous SSRP training among a school's teachers; (2) a "school management quality index" (see the appendix to our pre-analysis plan, which is in Appendix C); (3) teacher contract status (permanent contract or not); (4) teachers' years of teaching experience (whether they have over five years of experience); (5) student preparation and ability (see section 4.3 for the three ways we do this); and (6) student socio-economic characteristics, including gender, caste/ethnicity/religion groups, whether a parent has at least lower secondary education, and a family asset index. This asset index is constructed by applying IRT analysis to students' dichotomous answers to questions about whether their family owns a television, a bicycle, a scooter, a refrigerator and a computer and is highly correlated with a simple sum of the five asset indicators. The mean family owned 1.3 of these assets.

Standard errors*.* We cluster standard errors at the school level, because treatment was assigned at that level.[25]

Attrition*:* For school-level outcomes, there is no attrition because head teachers were interviewed in all sample schools at endline. For teacher-level outcomes, we focus on the treatment status of teachers at endline, and thus with endline teacher questionnaire response rates, rather than with attrition *per se.* This is because we believe that the best measure of *de facto* treatment in a school is whether the teachers in that school at endline were treated, regardless of their presence in that school at baseline, because these teachers are likely the relevant ones for the students during most of the treatment period (the academic year that was just ending when endline data were collected in early 2019). We have complete information (from monitoring data) on teacher's treatment status in the schools at endline, but teacher non-response is a potential issue for teacher-level outcomes measured in teacher questionnaires. These questions may lack responses for three reasons. First, interviewers were instructed not to administer the teacher questionnaire to the 6.9% of the teachers who were also the respondents for the head teacher questionnaire. Second, interviewers also failed to attempt interviews for another 2.7% of the teachers. Third, for another 6.9% of the teachers, interviewers could not administer the questionnaire due to teacher

[24] In Grade 9, there were 154 students with more than 1 math teacher (2.26%) and 281 with more than 1 science teacher (4.1%). In Grade 10, there were 137 students with more than 1 math teacher (2.4%) and 320 with more than 1 science teacher (5.5%). For these students we used average teacher characteristics.
[25] Our pre-analysis plan stated: "In specifications that include the pre-estimated measure of school management quality (see Appendix B), robustness will be assessed by calculating bootstrapped standard errors that account for this." Our standard errors below are very large, so this bootstrapping adds no value.

absence. All teachers who were present, and were asked to do so, completed the questionnaire. The difference between treatment and control samples in the rate of nonresponse for any reason is marginally significant (p-value=0.088), with a 5 percentage points lower rate in the treatment group, based on WLS teacher-level regressions controlling for district-stratum fixed effects, with standard errors adjusted for school-level clustering and stratification. This mostly reflects a larger fraction of teachers in the treatment group being respondents for the head teacher questionnaire.

For student-level outcomes, we experienced attrition of diverse sorts. Among 9th graders, who were in grade 8 at baseline, we can identify baseline students not present in the endline data for the first six reasons in Table 1. For 13 of our 203 schools (all 12 schools in Jumla district, and one in Panchthar district), grade 10 classes were not in session for the endline data collection visit, as students were released to study for the SEE. For these 13 schools, all grade 10 students, who were in grade 9 at baseline, are missing from the endline data (though many were still enrolled in school) and we cannot distinguish those missing due to classes not in session from those missing for other reasons (such as dropping out or moving to another school). For the other 190 schools, we observe attrition for the same six reasons recorded for Grade 9 students. As seen in Table 1, average academic performance varies over students who leave the sample for different reasons.

Table 1 shows that overall attrition rates, while high (36% in grade 8, 43% in grade 9), are similar for the treatment and control schools, differing by only 0.5 percentage points for grade 8 and 2.9 percentage points for grade 9. Also, attitors' average test scores are lower in the treatment groups than in the control group. Thus 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.[26]

Multiple hypothesis testing. We estimate impacts for many outcomes, so we may obtain at least some apparently significant impacts by chance. To account for this, we report False Discovery Rate adjusted p-values (henceforth q-values), for three sets of tests: (1) 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 for students with panel data, and separately for scores that use all test items and scores using only items most closely related to the SSDP training; (2) tests of no differences across actual treatment status for all LATE estimates of program impacts on test scores, also separately for estimates using all endline students and for students with panel data, and separately for scores that use all test items and scores using only items most closely related to the SSDP training; and (3) tests of no differences in outcomes across the treatment and control study arms for all ITT estimates of program

[26] Regression-based testing finds that attrition rates are insignificantly different between treatment and control groups for grade 8 baseline students, but significantly different for grade 9 baseline students. We also find statistically significant differences between treatment and control groups in the types of attrition and average test scores of attritters. Our pre-analysis plan calls for using Lee bounds of impact estimates if there are statistically significant differences in attrition across study arms, if there are differences of at least 5 percentage points in attrition rates across study arms, and if there are statistically significant or economically important differences in attrition types or average test scores among attritters. Following this rule requires us to calculate Lee bounds. Yet given the small differences in attrition between treatment and control, we do not believe Lee bounds calculations could alter our conclusions, and we do not report them.

Table 1: Description of student attrition

	Students in Grade 8 at Baseline			Students in Grade 9 at Baseline		
	% of all baseline students	Average Baseline Math Score[a]	Average Baseline Science Score[a]	% of all baseline students	Average Baseline Math Score[a]	Average Baseline Science Score[a]
Treatment Group						
All types of attrition	35.4	-0.33	-0.41	44.3	-0.20	-0.25
Enrolled in grade but absent	21.2	-0.33	-0.44	27.1	-0.21	-0.29
Repeating previous grade	1.9	-0.50	-0.58	2.3	-0.27	-0.20
Moved to other government school	3.1	-0.01	0.01	0.7	0.05	-0.39
Moved to private school	0.1	1.45	1.10	0.1	-0.79	-0.37
Not in school	7.9	-0.39	-0.45	7.2	-0.34	-0.36
Unknown	1.2	-0.48	-0.56	0.7	-0.54	-0.57
Classes not in session (school-level attrition)	0.0	NA	NA	6.3	0.03	0.07
Control Group						
All types of attrition	35.9	-0.21	-0.30	41.4	-0.19	-0.20
Enrolled in grade but absent	17.8	-0.16	-0.31	25.1	-0.05	-0.06
Repeating previous grade	4.7	-0.28	-0.31	2.5	-0.05	-0.14
Moved to other government school	3.1	0.04	0.03	0.4	-0.11	-0.11
Moved to private school	0.3	0.29	0.31	0.0	0.23	-0.20
Not in school	9.0	-0.34	-0.36	7.2	-0.33	-0.38
Unknown	1.1	-0.36	-0.54	0.4	-0.06	-0.55
Classes not in session (school-level attrition)	0.0	NA	NA	5.9	-0.70	-0.56
P-values for tests of treatment/control difference in:						
Attrition probability[b]		0.867			0.065	
Distribution of endline attrition status[c]		0.000			0.011	
Difference in average math score between attritters and stayers[d]		0.215			0.902	
Difference in average science score between attritters and stayers[d]		0.093			0.648	

Note: [a]Test scores are indices constructed from joint IRT analysis of baseline and endline scores for a given grade-level cohort and subject. [b]Test of hypothesis that coefficient on "treat" indicator is zero in a WLS student-level regression of an attrition indicator on the "treat" indicator and district-stratum fixed effects; standard errors are adjusted for stratification and school-level clustering. [c]Chi-squared test in unweighted data of hypothesis that distribution of students by endline presence or attrition type is identical for treatment and control. [d]Test of hypothesis that coefficient on interaction term between an indicator of subsequently attritting and the "treat" indicator is zero, in a WLS student-level regression of test score on attrition indicator, "treat" indicator, their interaction, and district-stratum fixed effects.

impact on teacher pedagogical practices from the head teacher, teacher and student questionnaires. We calculate q-values for each test, following the approach of Benjamini and Yekutieli (2001), which for allows any type of correlation across individual p-values within the group.

4. Findings

4.1 Intervention implementation fidelity

This section uses monitoring data (Shrestha 2019) and selected endline data to describe the roll-out of, and teacher attendance at, the SSDP training sessions. Section 4.3 uses a wider range of qualitative and quantitative data to check assumptions underlying the program's theory of change.

SSDP training roll-out and uptake. The SSDP trainings appear to have rolled out successfully in the study districts, though more slowly than we would have liked for the evaluation. Table 2 summarizes data on the timing, curriculum content of, and participation in, the SSDP math and science trainings for the 14 ETCs and 16 districts relevant to our study. The third and eighth columns show when the ETC math and science trainings were conducted. While the original evaluation plan was to roll out the trainings in October and November of 2017, soon after baseline data collection, in practice no trainings began before December of 2017, and most took place in April or May of 2018. This was often due to delays in earmarking and disbursing government funds by the Ministry of Finance and the Ministry of Education, Science and Technology. By the end of 2017 the Ministry of Finance had made available adequate funds to finance trainings in only one subject per district. We asked the ETCs to prioritize the math trainings, if feasible, in order to study the impacts of at least one common training across all study districts even if funding did not arrive for another set of trainings. Twelve ETCs complied, conducting the math training first, but the Dang and Myagdi ETCs conducted the science training first, because they failed to elicit enough applicants to the math training within the relevant time frame to warrant running that training first. Despite requests to the legislature by policymakers involved in this research to expedite funding for the remaining trainings, that funding was released only in mid-April of 2018.[27]

The ETCs seem to have adhered to the plan to invite treatment schools to send their secondary math and science teachers to the SSDP trainings, while not inviting teachers from control schools and other schools in the same small local areas (VDCs) as the control schools. Monitoring phone calls confirmed this, and administrative attendance data merged with our endline data indicate that endline teachers only in schools assigned to treatment attended the SSDP trainings.

[27] Additional idiosyncratic problems led to further delays in some ETCs: a delay in sending to the ETCs an NCED circular officially approving the unusual elements of the roll-out dictated by the research design; delays in sending the NCED training curriculum guidelines; teacher unavailability during the winter break in mountain areas; teacher unavailability during March 2018 end-of-year exams and the start of the next academic year in April of 2018; unavailability of ETC personnel to release funds or oversee trainings; delays because ETCs were already reserved for other trainings; and (later in 2018) uncertainty due to the likely dissolution of some ETCs due to larger government reforms.

Table 2: SSDP math and science training roll-out

ETC name	Study district name	ETC math training session dates from 12/2017 to 5/2018	Total number of math teachers attending the training[a]	Whether used SSDP or earlier curriculum as base (math)	No. of math teachers in endline schools assigned to SSDP treatment	% of endline math teachers who completed SSDP training[b]	ETC science training session dates from 12/2017 to 7/2018	Total number of science teachers attending the training[a]	Whether used SSDP or earlier curriculum as base (science)	No. of science teachers in endline schools assigned to SSDP treatment	% of endline science teachers who completed SSDP training[b]
Doti	Achham	5/3 – 5/12	18	SSDP	9	44	5/23 – 6/1	28	SSDP	8	38
Rupendehi	Arghakhanchi	4/6 – 4/15	24	SSDP	8	75	5/17 – 5/26	23	SSDP	7	57
Rupendehi	Kapilvastu	4/6 – 4/15	24	SSDP	9	56	5/17 – 5/26	23	SSDP	7	43
Myagdi	Baglung	5/6 – 5/15	25	SSDP	8	75	4/8 – 4/17	15	SSDP	4	25
Chitwan	Chitwan	4/1 – 4/10	25	SSDP[c]	12	33	5/21 – 5/30	25	SSDP[c]	12	25
Surkhet	Dailekh	1/5 – 1/14	14	Earlier[d]	9	67	6/13 – 6/22	22	Earlier	7	57
Kavre	Kavre	4/30 – 5/9	25	SSDP	7	71	5/23 – 6/1	24	SSDP	6	50
Kavre	Solukhumbu	4/30 – 5/9	25	SSDP	9	56	5/23 – 6/1	24	SSDP	9	22
Tanahu	Lamjung	3/28 – 4/6	25	SSDP	6	83	5/29 – 6/7	23	SSDP	6	67
Morang	Morang	3/7 – 3/16	20	SSDP	15	53	5/10 – 5/19	22	Earlier	16	25
Nuwakot	Nuwakot	2/2 – 2/11	24	SSDP	6	83	5/3 – 5/12	25	SSDP	6	50
Illam	Panchthar	2/6 – 2/15	19	SSDP[c]	6	50	6/18 – 6/27	23	Unclear[g]	8	50
Parsa	Parsa	3/22 – 3/31	20	SSDP	7	71	7/7 – 7/16	23	SSDP	8	50
Dang	Salyan	5/17 – 5/26	25	SSDP[e]	6	67	12/19 –12/28	21	Earlier	6	50
Dhanusa	Sindhuli	12/27 – 1/5	28	Earlier	8	63	5/6 – 5/15	18	Unclear [g]	8	50
Jumla	Jumla	4/8 – 4/17	25	No data	8	25[f]	4/29-5/7	No data	No data	7	43[f]
Total			366		133	60		339		125	42

Note: [a]Including teachers from both study and non-study schools and including teachers who did not complete all 10 days. [b]Teachers who signed up for training are counted as having completed the training unless we have attendance records showing that they attended fewer than 6 days. Among those for whom we have records, more than 90% completed all 10 days, so we imputed completion to those without records. [c]The ETC reported that it followed the NCED curriculum, but the training schedule it submitted was not specific enough to confirm. [d] The ETC reported that it referred to the hard copy of the curriculum provided to the trainers during the Master Training of Trainers. It was not informed of the new curriculum. Its training schedule lacks sessions for Arithmetic, Statistics, and Probability. [e]The ETC reports that it mostly followed the NCED curriculum but also included some general pedagogy subjects. [f]This total is likely to be an undercount, because the ETC provided incomplete information. [g]The training schedule the ETC submitted was not specific enough to confirm whether it followed the new curriculum.

All ETCs conducted trainings for the required 10 days, but the degree of adherence to the NCED curriculum guidelines varied across ETCs (Shrestha 2019), as seen in the fifth and tenth columns of Table 2. Most of the 13 ETCs for which we have data used variants of the SSDP curricula distributed by the NCED for both math and science, but two used a curriculum developed for secondary math teachers prior to the distribution of the SSDP guidelines, and at least three used pre-existing curricula for the science trainings, saying that they did not receive the new guidelines from the NCED before the trainings, or that they were not even aware that new training guidelines existed. In addition, the NCED guidelines allow ETCs to customize a small percentage of content based on local teacher requests. Thus, the content varied even among ETCs using the NCED curriculum. For math trainings, some ETCs added content in curriculum analysis, micro-teaching, student assessment, or the nature, historical background and importance of math. For science trainings, topics were added on the use of information and communication technology, curriculum analysis, student assessment, or child psychology and learning. Given SSDP designers' aim of employing a more uniform teacher training curriculum than had been used under the SSRP, the observed curriculum variation was probably greater than policymakers intended, though this is not surprising given the historically decentralized nature of teacher training in Nepal and the weak oversight of the training centers by central administrators (as discussed further in Section 4.3).

Standard practice is for ETCs to staff trainings using their own staff and contracted "roster trainers" (who often work elsewhere as school resource persons, head teachers, teachers, NGO staff, or college professors) to implement trainings. Among ETCs for which we have data, trainings were conducted by 0 to 5 staff trainers and 0 to 6 roster trainers. ETC staff tended to have skills for primary and pedagogy concerns, and roster trainers tended to have math and science expertise.

Participation in the SSDP training by invited teachers was discouragingly low (see seventh and twelfth columns, Table 2). Among grade 9 and 10 math and science teachers in the endline survey in schools assigned to SSDP training, only 60% of math teachers and 42% of science teachers had completed the trainings. Participation was low because some of those present at baseline and who participated in the trainings later left the schools (to be replaced by untrained teachers), and because some present at baseline (and endline) did not attend the trainings. Of the teachers present at endline in schools assigned to treatment and who did not complete the training, 60% were present at the baseline. Possible reasons for low participation are examined in Section 4.3.

On a more positive note, among teachers who did participate in the trainings, daily attendance was high, with nearly all teachers who enrolled in the trainings attending all 10 days.

4.2 Impact analysis

4.2.1 Descriptive statistics and balance tables

Tables 3 and 4 present descriptive statistics and balance tests for our baseline data. We see little evidence of imbalance; the last columns show significant differences in means between the treatment and control arms only for the number of students, the school having electricity, the head

Table 3: Baseline descriptive statistics and balance tests: schools and teachers

Variable	Number of observa-tions	Mean (standard error of mean)	Standard deviation	Mean for Treated Sample (std. dev.)	Mean for Control Sample (std. dev.)	p-value for test of $\beta_T = 0$[a,b,c]
School-level characteristics						
Total number of students in school	203	427.7 (15.4)	249.8	402.1 (214.5)	453.1 (277.7)	0.036**
Hours walking to nearest all-weather road	203	3.16 (0.35)	4.54	2.85 (3.64)	3.47 (5.26)	0.395
Students per section in Grade 9	203	49.53 (1.61)	26.74	47.93 (25.47)	51.11 (27.82)	0.242
Students per section in Grade 10	203	42.92 (1.38)	22.05	41.62 (20.73)	44.20 (23.19)	0.279
Days school was open last year (Grade 9)	203	195.59 (1.21)	16.18	194.87 (18.36)	196.31 (13.68)	0.544
School has electricity (several hours most days)	203	0.77 (0.03)	0.42	0.81 (0.40)	0.73 (0.44)	0.085*
Whether head teacher has at least Master's degree	203	0.57 (0.04)	0.50	0.63 (0.49)	0.52 (0.50)	0.078*
Hours per week head teacher teaches	203	16.38 (0.49)	6.82	16.95 (6.49)	15.81 (7.08)	0.193
Estimated management quality index	201	-0.02 (0.07)	0.90	0.04 (0.90)	-0.07 (0.89)	0.454
Teacher-level characteristics						
Is female	395	0.07 (0.01)	0.25	0.09 (0.29)	0.04 (0.20)	0.030**
Has at least Bachelor's degree in math/science	361	0.81 (0.02)	0.40	0.79 (0.41)	0.82 (0.38)	0.291
Had SSRP training	395	0.31 (0.03)	0.46	0.30 (0.46)	0.32 (0.47)	0.481
Years of experience	393	11.09 (0.44)	8.12	10.69 (7.80)	11.49 (8.42)	0.168
Hours spent preparing for class	393	0.81 (0.05)	0.96	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 treatmentindicator and district and priority stratum fixed effects. [c]For binary outcome variables, weighted probit regressions were used instead of WLS.

Table 4: Baseline descriptive statistics and balance tests: students

Variable	Number of observa-tions	Mean (standard error of mean)	Standard deviation	Mean for Treated Sample (std. dev.)	Mean for Control Sample (std. dev.)	p-value for test of $\beta_T = 0$[a.b.c]
Grade 8 and 9 student-level characteristics						
Is female	16435	0.55 (0.01)	0.50	0.55 (0.50)	0.55 (0.50)	0.875
Father can read and write	15594	0.83 (0.01)	0.38	0.83 (0.37)	0.82 (0.38)	0.470
Father has at least secondary education	15753	0.28 (0.01)	0.45	0.28 (0.45)	0.28 (0.45)	0.881
Mother can read and write	14830	0.59 (0.01)	0.49	0.59 (0.49)	0.59 (0.49)	0.964
Mother has at least secondary education	15831	0.11 (0.01)	0.32	0.11 (0.31)	0.12 (0.33)	0.384
Nepalese is main language spoken at home	16251	0.75 (0.02)	0.43	0.74 (0.44)	0.77 (0.43)	0.226
Family IRT asset index[d]	16435	-0.04 (0.03)	0.77	-0.03 (0.77)	-0.04 (0.77)	0.950
Baseline test scores						
Grade 8 math percentage score	7651	18.12 (0.45)	10.57	18.13 (10.52)	18.11 (10.62)	0.460
Grade 8 math IRT latent variable	7651	0.01 (0.04)	0.86	0.01 (0.86)	0.01 (0.87)	0.414
Grade 8 science percentage score	7651	28.17 (0.68)	12.4	28.58 (13.07)	27.76 (11.66)	0.609
Grade 8 science IRT latent variable	7651	0.05 (0.05)	0.91	0.08 (0.95)	0.03 (0.86)	0.477
Grade 9 math percentage score	8784	28.53 (0.83)	15.64	27.95 (15.00)	29.06 (16.18)	0.168
Grade 9 math IRT latent variable	8784	0.02 (0.05)	0.92	-0.01 (0.88)	0.04 (0.95)	0.196
Grade 9 science percentage score	8784	28.07 (0.58)	11.11	27.61 (10.61)	28.48 (11.54)	0.314
Grade 9 science IRT latent variable	8784	0.02 (0.04)	0.84	-0.01 (0.81)	0.04 (0.88)	0.366

Note: [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.

teacher having at least a Master's degree, and whether a teacher is female. The absolute sizes of the differences are not large, and the differences lose significance after adjusting for multiple hypothesis testing. Below we check the robustness of our main findings to including controls for these variables. Following our pre-analysis plan, we also check robustness by including controls for key school-, teacher- and student-level variables of interest for studying impact heterogeneity.

Two dimensions of heterogeneity of interest to policymakers in Nepal are student gender and ethnicity. To assess heterogeneity in academic performance on these dimensions, we use endline rather than baseline test scores as we believe that the quality of our assessments and our ethnicity indicators is higher in the endline data. Average scores for females are lower than for males on both math and science for both grades. Muslim and Dalit students have lower scores than those from other ethnic groups, while Brahmin, Chhetri and Newar students fare better. This disparity is largest for grade 10 students in science. For details see Supplementary Table 1 in Appendix F.

4.2.2 Estimated impacts on student test scores

Table 5 presents descriptive statistics and ITT impact estimates for the eight endline student test score variables: two grades (9 and 10), two subjects (math and science), and two types of scores for each assessment: (1) a total score based on answers to all 35 questions in the assessment, and (2) an "SSDP focus" score based on answers to about half of the questions that we deemed most closely tied to the SSDP training curricula. All scores were normalized by subtracting the mean and dividing by the standard deviation of the control group.[28]

Our main ITT regressions using the full endline sample (Table 5, fourth column) yield no evidence that the SSDP teacher trainings increased student test scores. Five of the eight estimates are negative, and one is statistically significant at the 5% level while three others are significant at the 10% level. The 95 percent confidence intervals (not shown) rule out effects above 0.10 standard deviations (of the distribution of test scores) in three 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 5, seventh column), largely confirm these results. Baseline scores are highly significant, as expected, raising the R-squared measures. Adding them reduces slightly the standard errors of the estimates, ruling out positive effects at slightly lower thresholds.

ITT estimates may be small because only 60% (42%) of endline math (science) teachers in the schools assigned to the SSDP training actually attended those trainings. We report LATE in Table 6; these regressions include an indicator of whether students' teachers completed the SSDP trainings and use treatment assignment to instrument that variable. As expected, the LATE estimates of the impact of completion of treatment (rather than invitation to treatment), are larger in absolute value and less precise but tell largely the same story as the ITT estimates: almost all

[28] For most of the regressions in Tables 5 and 6, when we run similar regressions that allow different impacts between SSDP training with and without the video assignment, we fail to reject the hypothesis that the two impacts are equal. The only exception is grade 9 math, where the difference is significant at the 10% level for the total score and the 5 percent level for the SSDP focus score. After adjusting the p-values of these tests for multiple hypothesis testing, these differences in impacts for the two training types are insignificant.

Table 5: 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	R^2		Treat	Baseline Test Score	R^2	
Full assessments									
Grade 9 math	-0.057 (0.931)	0.000 (1.000)	-0.110* (0.066)	0.229	6,800	-0.107** (0.050)	0.532*** (0.020)	0.428	4,903
Grade 9 science	-0.051 (0.912)	0.000 (1.000)	-0.109* (0.060)	0.160	6,797	-0.106* (0.054)	0.494*** (0.022)	0.350	4,901
Grade 10 math	-0.035 (0.998)	0.000 (1.000)	-0.044 (0.072)	0.253	5,832	-0.000 (0.050)	0.563*** (0.017)	0.494	4,992
Grade 10 science	-0.002 (0.971)	0.000 (1.000)	0.006 (0.074)	0.181	5,829	0.025 (0.061)	0.502*** (0.021)	0.359	4,990
SSDP focus items									
Grade 9 math	-0.004 (0.953)	0.000 (1.000)	-0.046 (0.066)	0.163	6,800	-0.054 (0.056)	0.444*** (0.021)	0.307	4,903
Grade 9 science	-0.061 (0.914)	0.000 (1.000)	-0.100* (0.057)	0.126	6,797	-0.075 (0.054)	0.426*** (0.020)	0.254	4,901
Grade 10 math	-0.044 (1.001)	0.000 (1.000)	-0.037 (0.070)	0.207	5,832	-0.024 (0.059)	0.444*** (0.018)	0.352	4,992
Grade 10 science	0.009 (0.979)	0.000 (1.000)	0.024 (0.072)	0.142	5,829	0.022 (0.063)	0.433*** (0.019)	0.267	4,990

Notes: 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. Standard errors, in parentheses, account for random assignment within strata and are clustered at the school level. Statistical significance at .10, .05 and .01 levels indicated by *, ** and ***.

Table 6: IV/LATE estimates of impact of SSDP training on students' normalized test scores, full endline and panel samples

	Full Endline Sample Treatment Effect		Sample Size	Panel Sample Treatment Effect			Sample Size
	Treat	R^2		Treat	Baseline Score	R^2	
Full assessments							
Grade 9 math	-0.143 (0.094)	0.240	6,048	-0.145** (0.069)	0.526*** (0.022)	0.433	4,353
Grade 9 science	-0.279 (0.178)	0.145	5,963	-0.298* (0.157)	0.516*** (0.023)	0.350	4,336
Grade 10 math	-0.100 (0.105)	0.262	5,098	-0.023 (0.080)	0.537*** (0.020)	0.476	4,343
Grade 10 science	-0.050 (0.155)	0.175	5,003	-0.026 (0.122)	0.501*** (0.021)	0.363	4,310
SSDP focus items							
Grade 9 math	-0.046 (0.093)	0.177	6,048	-0.051 (0.076)	0.442*** (0.021)	0.315	4,353
Grade 9 science	-0.252 (0.165)	0.117	5,963	-0.167 (0.153)	0.445*** (0.021)	0.262	4,336
Grade 10 math	-0.100 (0.105)	0.263	5,098	-0.055 (0.091)	0.412*** (0.019)	0.336	4,343
Grade 10 science	-0.037 (0.148)	0.137	5,003	-0.048 (0.125)	0.433*** (0.020)	0.266	4,310

Note: All estimates are for β_T from instrumental variable (IV) regressions of normalized student assessment scores on a variable indicating that the student's teacher received training (instrumented by treatment assignment), district by priority stratum fixed effects, and dummy variables for whether assent was requested before or after the test and for whether the math test was administered first (followed by the science test). Regression standard errors, in parentheses, account for random assignment within strata and are clustered at the school level. Estimates statistically significant at the .10, .05 and .01 levels are indicated by *, ** and ***, respectively.

point estimates are negative and 14 of 16 are statistically insignificant. Note that LATE impact estimates apply only to teachers who participated in the training, and may not extrapolate well to those who did not participate. However, among the teachers in the treated schools when the endline data were collected, there are few observable differences between those who participated in the training and those who did not. This is shown in Online Appendix Table A2. (The only significant difference is that teachers who had received other training in the past were much more likely to participate in the SSDP training; 73% of participated teachers had had other training, but only 40% of those who did not participate.) Thus it is possible that these estimates apply more broadly to all teachers.

Following our pre-analysis plan, we checked our results' robustness in five ways. The results are very similar if we: (1) use unweighted instead of weighted regressions; (2) omit controls for test-taking conditions; (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 the dependent variable. The details are in Supplementary Table 2 in Appendix F.[29]

The descriptive statistics and results motivate two comments on our sample size calculations. First, the intra-school correlations (in the control group) ranged from 0.054 to 0.095, which is much lower than the value of 0.65 that we assumed for our power calculations (based on data from earlier assessments). Yet our standard errors are not much smaller than the sizes targeted in our minimum detectable effect (MDE) calculations. The main explanation is the lower-than-assumed treatment rate among treatment schoolteachers. Second, while our standard errors are about 0.07 standard deviations, and thus small enough to detect a true effect size of 0.14 standard deviations or larger (a reasonable MDE size for standard sample size calculations), they are still too large to rule out modest effects, such as 0.10 standard deviations, even for true effect sizes of zero.[30] This suggests that the MDE criteria often employed in sample size calculations will lead researchers to set samples that are too small to provide confident conclusions of no impact.

4.2.3 Analysis of impact heterogeneity

We comply with our pre-analysis plan to analyse the heterogeneity of (ITT) impacts along diverse dimensions, despite detecting no positive average impacts, since it is possible to detect positive impacts in some subsets of the population. We do not report heterogeneity results for our LATE estimates, since none of the 12 estimates for LATE heterogeneity was significant at the 5% level. This is not surprising given that we find little evidence of heterogeneity in our ITT estimates.

Regressions reported in Online Appendix A (Table A13) allow treatment effect heterogeneity by teacher and school characteristics, including whether: 1) the teacher was trained under the SSRP (the most recent previous government teacher training); 2) the teacher has a permanent contract; 3) the teacher has five years of experience or less; and 4) the school's estimated management quality index is above the median. Only one of the 16 interaction terms is statistically significant at the 5% level (another is significant at the 10% level), and in only two of 16 do we reject the null of no joint effect. The interaction terms' standard errors are large. We conclude that our data cannot detect impact heterogeneity in these dimensions.

[29] Following our pre-analysis plan, we also checked robustness using another measure of student achievement: scores of the School Education Examination (SEE). We checked balance according to these scores using the SEE results from shortly after the baseline data were collected (in March/April of 2018), finding scores to be marginally significantly lower in treatment schools. Because of this we used the SEE data from March/April 2018 and March/April 2019 to obtain difference-in-differences estimates (with school fixed effects) of the treatment impact on math and science SEE scores. The estimates are statistically insignificant and more imprecisely estimated than our main estimates. The point estimate for the SEE math school is similar to ours, while the point estimate for science is positive, yet not significant.

[30] A 95 percent confidence interval an unbiased estimate of a zero impact would range from -0.14 to 0.14 and thus would not rule out a true effect size of 0.10 standard deviations.

Table 7: Analysis of ITT combined treatment impact heterogeneity by student and household characteristics, full endline sample

	Grade 9		Grade 10	
	Mathematics	Science	Mathematics	Science
Heterogeneity by gender of student				
Treat	-0.142*	-0.161**	-0.142*	-0.071
	(0.068)	(0.062)	(0.081)	(0.088)
Female student	-0.329***	-0.339***	-0.475***	-0.486***
	(0.049)	(0.046)	(0.044)	(0.044)
Treat × Female student	0.057	0.093	0.174**	0.138*
	(0.063)	(0.059)	(0.072)	(0.075)
P-value for test of joint significance of Treat and interaction term	0.116	0.028**	0.045**	0.184
R^2	0.253	0.183	0.292	0.227
Sample size	6,801	6,798	5,833	5,829
By education of parents				
Treat	-0.090	-0.059	-0.035	0.023
	(0.075)	(0.072)	(0.073)	(0.079)
At least one parent had secondary Education	0.327***	0.322***	0.254***	0.235***
	(0.041)	(0.047)	(0.046)	(0.056)
Treat × At least one parent had secondary education	-0.035	-0.093	-0.025	-0.041
	(0.065)	(0.071)	(0.064)	(0.081)
P-value for test of joint significance of Treat and interaction term	0.177	0.052*	0.756	0.877
R^2	0.253	0.181	0.266	0.192
Sample size	6,801	6,798	5,833	5,829
By household wealth				
Treat	-0.111*	-0.112*	-0.032	0.016
	(0.062)	(0.057)	(0.070)	(0.074)
Household wealth index	0.202***	0.194***	0.161***	0.116**
	(0.036)	(0.038)	(0.046)	(0.047)
Treat × Household wealth index	-0.004	-0.076	0.021	0.048
	(0.056)	(0.053)	(0.064)	(0.071)
P-value for test of joint significance of Treat and interaction term	0.206	0.039**	0.826	0.793
R^2	0.248	0.172	0.264	0.188
Sample size	6,739	6,736	5,769	5,763

Note: All estimates are for β_T from WLS regressions of normalized student assessment scores on the treatment indicator, the heterogeneity variable, the interaction of the treatment indicator and heterogeneity variable, district by priority stratum fixed effects, and dummy variables for whether assent was requested before or after the test and for whether the math test was administered first (followed by the science test). Regression standard errors, in parentheses, account for random assignment within strata and are clustered at the school level. Estimates statistically significant at the .10, .05 and .01 levels are indicated by *, ** and ***, respectively.

The regressions in Table 7 allow for heterogeneity in treatment effects by student and household characteristics.[31] The variables themselves – whether the student is female, whether at least one parent has secondary education, and a household wealth index – are significantly associated with test scores in the expected directions. Yet only two of the 12 interactions are statistically significant, again in part due large standard errors. In just four of 12 cases are the treatment indicator and the interaction term jointly significant. Thus, while the point estimates suggest that the training's impact was less negative for girls than for boys in grade 9, positive for girls while negative for boys in grade 10, and slightly more negative for students with more educated parents, the statistical imprecision of these estimates prevents any strong conclusions to be drawn. Online Appendix A (Table A14) shows similarly inconclusive heterogeneity results by student ethnicity.

We investigated whether the impacts differ by student academic ability using two indicators of the academic achievement they brought to secondary school: tercile in the baseline test score distribution and percentage correct on the endline assessment items pertaining to knowledge from earlier grades. (For the latter, the outcomes of interest are students' scores only on the endline assessment items that map to grade 9 and 10 curriculum.) Online Appendix A (Table A15) shows that these interaction terms are all statistically insignificant at the 5% level (2 of 12 are significant at the 10% level), in part due to large standard errors. Table 8 shows the results of quantile regressions that together describe the treatment impacts at the 10th, 25th, 50th, 75th and 90th percentiles of the endline test score distribution. Our aim is to differentiate impacts across students from the 10th to 90th percentiles of the *unconditional* endline performance distribution, rather than students at those percentiles of the endline performance distribution *conditional on* their district, stratum, or student or school characteristics. Thus these quantile regressions have no controls other than the treatment indicator.[32] In Table 8's top panel, the standard errors account for clustering at the school level, but the estimation is not weighted (given limitations in STATA). Estimates in the bottom panel are weighted, but the standard errors do not account for clustering. The bottom panel confirms that weighting (or lack thereof) has little effect on the point estimates, so we focus on the top panel, which has clustered standard errors.[33] The point estimates suggest some tendency for any substantial negative impacts of SSDP teacher training to be concentrated among students at the upper end of the performance distribution, though again, standard errors (especially those adjusted for clustering) are large.

A tendency for negative impacts to be concentrated among students who performed best prior to the intervention, which is weakly suggested by our results, is plausible. The teaching innovations

[31] Analogous tables with separate effects by gender, parental education and teacher contract status are in Appendix Table A16. These are for researchers who want to conduct a meta-analysis using our results.

[32] In our pre-analysis plan we proposed to estimate these impacts using the generalized quantile regression method proposed by Powell (2017), which estimates impacts on unconditional quantiles while allowing for controls. Having encountered computational problems, we concluded that this method is infeasible for this report. Given that treatment was randomly assigned, we do not require other regressors for unbiasedness. Without additional controls, standard quantile regressions estimate the impacts of the treatment on quantiles of the unconditional distribution.

[33] For other regression results that allow for both weighting and clustered standard errors, we find that using population weights has relatively little effect on the results, while clustering the standard errors has a large negative effect on the precision of the estimates. Note that the results with clustered standard errors were estimated using a block (cluster) bootstrap with 50 replications.

Table 8: Analysis of ITT combined treatment impact on quantiles of the endline test score distribution

	Grade 9		Grade 10	
	Mathematics	Science	Mathematics	Science
Clustered standard errors (via bootstrapping), no weights				
Treat, quantile 0.10	0.020	0.051	0.032	0.031
	(0.086)	(0.081)	(0.079)	(0.077)
Treat, quantile 0.25	0.046	0.042	-0.021	0.021
	(0.069)	(0.064)	(0.085)	(0.088)
Treat, quantile 0.50	-0.019	-0.010	-0.116	-0.051
	(0.087)	(0.076)	(0.131)	(0.122)
Treat, quantile 0.75	-0.125	-0.088	-0.140	-0.129
	(0.111)	(0.104)	(0.162)	(0.148)
Treat, quantile 0.90	-0.150	-0.232**	-0.184	-0.169
	(0.116)	(0.098)	(0.141)	(0.127)
Weighted estimation, standard errors not adjusted for clustering				
Treat, quantile 0.10	-0.010	-0.009	0.002	0.041
	(0.039)	(0.038)	(0.039)	(0.042)
Treat, quantile 0.25	0.011	0.017	-0.033	0.025
	(0.033)	(0.032)	(0.033)	(0.036)
Treat, quantile 0.50	-0.035	-0.049	-0.083**	0.016
	(0.035)	(0.033)	(0.040)	(0.036)
Treat, quantile 0.75	-0.116***	-0.085**	-0.050	-0.041
	(0.040)	(0.040)	(0.050)	(0.050)
Treat, quantile 0.90	-0.122**	-0.210***	-0.060	-0.048
	(0.050)	(0.053)	(0.066)	(0.052)
Sample size	6,801	6,798	5,833	5,829

Note: All estimates are for β_T from quantile regressions of normalized student assessment scores on the treatment indicator. The standard errors are in parentheses; in the top panel they account for random assignment within strata and are clustered at the school level, but are not weighted. In the bottom panel the standard errors are not adjusted for strata or clustering, but they incorporate school-level population weights. Estimates statistically significant at the .10, .05 and .01 levels are indicated by *, ** and ***, respectively.

promoted by the SSDP training (requiring demonstrations relevant to advanced topics using teaching aids made from local materials) may have been largely irrelevant for students entering grades 9 and 10 with below-grade-level subject knowledge (leaving them with little impact) while having negative impacts on students who had done better prior to their teachers' SSDP training, since, for example, the demonstration-based methods may have slowed teachers down, causing them to cover fewer curriculum concepts or fewer examples for any one concept.

4.2.4 Estimated impacts on intermediate outcomes

Our aim in this evaluation was not just to estimate, but also to understand, the impacts of SSDP teacher training on student test scores. Having concluded that the trainings failed to raise student test scores, we now ask why this was so. In this section we examine the program's impact on

Table 9: Summary of ITT estimates of impact on teacher subject knowledge and attitude

Intermediate Outcomes	Sample Size	Descriptive Statistics		Estimation method for testing	p-value of test of no impact
		Control	Treatment		
Teacher subject knowledge assessment scores					
Math (mean/std. dev.)	246	0.000 (1.000)	-0.014 (1.010)	WLS	0.907
Science (mean/std. dev.)	233	0.000 (1.000)	-0.136 (0.966)	WLS	0.298
Head teacher reports on teacher skill and interest					
Whether teacher is very interested in learning ways to teach more effectively (%)	443	46.9	45.0	Probit	0.736
Teacher's command of subject matter (% distribution)	442	--	--	Ordered Probit	0.566
Very weak or partial	--	11.7	11.0	--	--
Good	--	67.8	70.6	--	--
Excellent	--	20.5	18.3	--	--

Note: The unit of observation is the teacher. For continuous outcome variables, descriptive statistics columns report weighted means and standard deviations. For dichotomous outcomes, they report the weighted percentage for which the outcome is true. For ordered polychotomous variables, they report the weighted percentage distributions by category. Tests of no impact of the SSDP teacher training are based on regressions of the dependent variable on the treat variable and strata dummy variables, using WLS estimation for continuous variables, 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. Teacher assessment scores are derived from "Student Assessment Item Evaluations Forms" that teachers were asked to complete anonymously (see text).

outcomes related to teachers' subject knowledge and teaching practices. Section 4.3 below examines in greater depth the assumptions underlying the theory of change that might be violated, in ways that help to explain the disappointing lack of impact on intermediate and final outcomes.

Table 9 reveals little or no impact of the SSDP trainings on teacher subject knowledge or attitudes; Online Appendix A (Tables A17-A19) gives regression details. Aside from the teacher science evaluation outcome (where the treatment group mean is *lower* than the control group mean), the descriptive statistics are very close for treatments and controls.

Taking a broad view of teaching practices, we first examine impacts on teacher attendance at school. For our three measures of teacher attendance, we again detect no impact, as shown in Table 10. The measures come from direct observation by enumerators, head teacher reports and student reports. Descriptive statistics are very similar for treatment and control groups, suggesting that the failure to reject the null of no impact is not simply the result of imprecise estimation.

Tables 11, 12 and 13 present results for teaching practice indicators derived from head teacher, teacher and student questionnaires. (Details are in Online Appendix A Tables A23-A39). We find very few statistically significant differences in teaching practices between the treatment and

Table 10: Summary of ITT estimates of impacts on teacher attendance

Intermediate outcome	Sample Size	Descriptive Statistics		Estimation Method for Testing	p-value for Test of No Impact
		Control	Treatment		
Teacher is present on 1st day of school visit (enumerator observation) (%)	434	91.5	90.5	Probit	0.422
Regularity of teacher attendance (reported by head teacher) (%)	443			Ordered Probit	0.167
90% or higher		59.7	57.7		
80-89%		32.3	27.6		
Less than 80%		8.0	14.7		
Frequency of math teacher absence (report by student) (%)	12,602			Ordered Probit	0.850
Never absent		23.6	21.0		
Absent 1 to 2 times per month or less		55.4	58.3		
Absent 3 or 4 times per month/ at least once per week		21.0	20.6		
Frequency of science teacher absence (report by student) (%)	12,573			Ordered Probit	0.144
Never absent		22.4	18.7		
Absent 1 to 2 times per month or less		54.1	55.5		
Absent 3 or 4 times per month/at least once per week		23.5	25.8		

Note: The unit of observation is the teacher. 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. Tests of no impact of the SSDP teacher training are based on regressions of the dependent variable on the treat variable 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. Note that the probit regression is based on only 333 observations because adjusting the standard errors for stratification drops strata that have only one observation.

control groups, yet two of the three exceptions – head teacher reports on teacher's frequency of using teaching materials or visual aids and teacher's 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 one other exception, whether a teacher required students to work on longer-term projects is not obviously related to SSDP training. One other statistically significant difference is in the unexpected direction, with treatment group teachers *less* likely to use written lesson plans. Multiple hypothesis testing adjustments eliminate statistically significant differences. Comparisons of descriptive statistics for all these outcomes suggest, though, that even if some of differences are statistically significant, they are not very large, with the probabilities of teachers employing practices or employing them frequently different between treatment and control groups by only a few percentage points.

Table 11: ITT estimates of impacts on math teacher teaching practices (student reports)

Intermediate outcome	Sample Size	Descriptive Statistics		Estimation Method for Testing	Test of No Impact (p-value)
		Control	Treatment		
Gives homework all days (%)	12,530	80.4	80.6	Probit	0.960
Homework checking frequency (% distribution):	12,482	--	--	Ordered Probit	0.522
Up to once a week week		14.8	15.4	--	--
2-3 times a week		32.1	33.9	--	--
Every day		53.1	50.6	--	--
Homework correction frequency (% distribution):	12,505	--	--	Ordered Probit	0.925
Up to once a week		13.9	15.1	--	--
2-3 times a week		27.9	26.1	--	--
Every day		58.2	58.8	--	--
Interactive teaching frequency (% distribution):	12,535	--	--	Ordered Probit	0.272
Up to once a week		14.0	16.8	--	--
2-3 times a week		32.7	32.7	--	--
Every day		53.3	50.5	--	--
Class time group work frequency (% distribution)	12,572	--	--	Ordered Probit	0.755
Never		32.5	31.4	--	--
Less than once a week		7.6	6.8	--	--
Once a week		16.9	20.5	--	--
2-3 times a week		26.6	27.8	--	--
Every day		16.4	13.4		
Frequency of using local materials or visual aids (% distribution):	12,552	--	--	Ordered Probit	0.709
Never		26.0	24.1	--	--
Less than once a week		11.9	11.5	--	--
Once a week		18.4	24.4	--	--
2-3 times a week		24.6	24.1	--	--
Every day		19.2	15.9	--	--
Frequency of using materials from internet (% distribution):	12,574	--	--	Ordered Probit	0.904
Never		32.2	33.4	--	--
Less than once a week		9.9	9.4	--	--
Once a week		18.4	21.4	--	--
2-3 times a week		23.3	23.0	--	--
Every day		16.2	12.8		

Note: 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. Tests of no impact of the SSDP teacher training are based on regressions of the dependent variable on the treat variable 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 12: ITT estimates of impacts on science teacher teaching practices (student reports)

Intermediate outcome	Sample Size	Descriptive Statistics		Estimation Method for Testing	Test of No Impact (p-value)
		Control	Treatment		
Gives homework frequency (% distribution)	12,532	--	--	Ordered Probit	0.793
Up to once a week		14.9	16.0	--	--
2-3 times a week		33.0	32.0	--	--
Every day		52.1	52.0	--	--
Homework checking frequency (% distribution):	12,491	--	--	Ordered Probit	0.639
Never		4.3	3.7		
Up to once a week		19.7	22.3	--	--
2-3 times a week		35.7	36.2	--	--
Every day		40.3	37.8	--	--
Homework correction frequency (% distribution)	12,571	--	--	Ordered Probit	0.453
Never		4.6	4.9		
Up to once a week		16.1	17.7	--	--
2-3 times a week		33.7	33.5	--	--
Every day		45.6	43.9	--	--
Interactive teaching frequency (% distribution):	12,550	--	--	Ordered Probit	0.519
Up to once a week		18.1	18.7	--	--
2-3 times a week		35.9	37.2	--	--
Every day		46.0	44.1	--	--
Class time group work frequency (% distribution)	12,559	--	--	Ordered Probit	0.985
Never		26.9	25.7	--	--
Less than once a week		7.4	7.5	--	--
Once a week		19.1	21.7	--	--
2-3 times a week		28.3	29.5	--	--
Every day		18.2	15.7		
Frequency of using local materials or visual aids (% distribution):	12,574	--	--	Ordered Probit	0.951
Never		18.7	17.0	--	--
Less than once a week		10.1	10.7	--	--
Once a week		21.8	24.8	--	--
2-3 times a week		29.3	29.8	--	--
Every day		20.1	17.8	--	--
Frequency of using materials from the internet (% distribution):	12,561	--	--	Ordered Probit	0.391
Never		23.5	25.3	--	--
Less than once a week		9.9	10.7	--	--
Once a week		20.4	22.1	--	--
2-3 times a week		29.6	28.6	--	--
Every day		16.5	13.3		

Note: See notes to Table 11.

Table 13: ITT estimates of impacts on teaching practices: head teacher and teacher reports

Intermediate outcome	Sample Size	Descriptive Statistics		Estimation Method for Testing	Test of No Impact (p-value)
		Control	Treatment		
Head Teacher reports					
Teacher ever creates teaching materials from local resources (%)	437	61.0	59.0	Probit	0.960
Teacher's frequency of using teaching materials or visual aids (%distribution)	438	--	--	Ordered Probit	0.072*
Never		21.5	16.8	--	--
Sometimes (less than once per week)		69.5	65.2	--	--
Often (one or more times per week)		8.9	18.0	--	--
Teacher ever collects or requires students to collect local information (%)	422	35.4	35.2	Probit	0.816
Teacher frequency requiring students to work in small groups (% distribution)	441	--	--	Ordered Probit	0.079*
Never		19.6	13.7	--	--
Sometimes (less than once per week)		67.5	67.0	--	--
Often (one or more times per week)		13.0	19.2	--	--
Teacher ever requires students to work on longer term projects (%)	439	50.7	58.6	Probit	0.013**
Teacher self-reports					
Preparation per class session (minutes)	401	32.3 (2.9)	30.4 (2.6)	WLS	0.423
Ever uses written lesson plan (%)	401	12.0	6.3	Probit	0.101
Frequency of requiring students to work in small groups (% distribution):	401	--	--	Ordered Probit	0.953
Never		15.3	13.0	--	--
Less than once a week		28.7	23.7	--	--
Once a week		30.5	39.8	--	--
Two or more times a week		25.4	23.5	--	--
Frequency of using classroom examples/homework involving local information (% distribution):	401	--	--	Ordered Probit	0.478
Less than once a month or never		24.2	20.3	--	--
Once a month		26.0	28.8	--	--
Less than once a week but more than once a month		4.9	12.3	--	--
Once a week		25.7	23.0	--	--
Two or more times a week		19.2	15.8	--	--
Frequency of requiring students to collect local information (% distribution)	401	--	--	Ordered Probit	0.561
Never		36.4	29.9	--	--
Less than once a month		16.7	19.0	--	--
Once a month		21.3	26.3	--	--
Less than once a week but more than once a month		7.1	6.8	--	--
Once a week or more		18.6	17.9	--	--

Note: The unit of observation is the teacher. For continuous outcome variables, the descriptive statistics

are weighted means and standard deviations. For dichotomous outcomes, the descriptive statistics are weighted percentages for which the outcome is true. For ordered polychotomous variables, they report the weighted percentage distributions by category. Tests of no impact of the SSDP teacher training are based on regressions of the dependent variable on the treat variable 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. Note that the probit regressions use fewer observations because adjusting the standard errors for stratification drops strata that have only one observation; see online appendix tables for those sample sizes.

4.3 Violations of the assumptions underlying the SSDP training program's theory of change

This section examines the assumptions underlying the SSDP teacher training program's theory of change (see Figure 1), seeking to understand why the trainings failed to raise 9th and 10th grade students' math and science test scores. We examine potential barriers to desirable impact by three key sets of actors along the logical chain connecting the teacher training policy idea to the ultimate desired impact on student learning: (1) trainers, who must exercise skill and care in delivering the training curriculum; (2) teachers, who must attend the trainings, learn from them, and then implement new or improved teaching methods in their classrooms and (3) students, who must be prepared to learn more when teaching improves. For each of these groups we ask: Under what conditions would this group respond to the policy mandate or invitation in ways that lead to strong, positive impacts on learning? More specifically, what must be true of the actors' capacity (knowledge, skills, guidance and decision-making scope), the resources provided to them, and their motivation (either intrinsic or induced through external accountability) for them to respond well? Which of these conditions seem to have been fulfilled well, and which may not have been fulfilled well? We draw on four complementary types of data (often finding agreement across multiple sources): (1) monitoring data gathered directly from ETCs via research team phone calls (Shrestha 2019); (2) small-N study using in-depth in-person interviews (Acharya and Uprety 2019); (3) telephone survey with trainers and teachers who attended SSDP trainings (Schaffner, Glewwe and Sharma 2019a); and (4) data from the baseline and endline quantitative surveys.

Trainers. For SSDP training to succeed, trainers must have adequate knowledge and skills for following any directions given to them, and for developing and delivering high quality training content. In addition, they must have adequate time for preparation, adequate materials and facilities, and adequate motivation to work hard toward the objective of helping teachers improve their teaching practices. The motivation may be internal motivation, arising out of their own desire to do a good job or to advance educational development, or it may be external motivation, arising out of the knowledge that they will be held accountable for their performance, in the sense that they will receive (monetary or non-monetary) rewards if they do well or penalties if they do poorly.

Many trainers seem to have executed SSDP trainings with at least a moderate level of good will. According to administrative and phone interview data, all the trainings in our study districts lasted the mandated 10 days, and trainers developed training materials laid out in more detail than in the documents provided by the NCED, suggesting that they spent some time preparing. In phone interviews, trainers expressed clearly the changes in teaching practices they wanted participants to make, and all but one had suggestions for improving the trainings (demonstrating engagement

with their assignments). Telephone interview data also suggest that trainers' instruction methods were reasonably interactive. Three quarters (74%) of teachers mentioned group problem solving as the first or second largest use of time during the ETC sessions. Using such methods may be beneficial not only to help teachers learn other methods, but also as an example for teachers to follow. (We know from baseline data that teachers typically make little use of group work in their classes, spending most of their teaching time lecturing from the blackboard.[34])

The data also suggest, however, that the trainings' quality fell significantly short of the ideal. When trainers were asked in telephone interviews to describe the guidance and training they had been given by the NCED to prepare them for the trainings, all the interviewed trainers responded that they had received a syllabus for the training to guide their preparation, but 19 out of 23 mentioned that they received no training of trainers for the SSDP trainings (though they had received training of trainers for previous waves of training). Many mentioned lack of detail in the guidelines for the trainings and that they received the syllabus with too little time to prepare for the training sessions. In open-ended responses, teachers reported that trainers seemed inadequately prepared.[35] In addition, some trainers' attitudes may have diminished their training efforts: 61% mentioned inadequate teacher motivation, poor teacher attitudes and/or lack of monitoring of teachers when asked what main 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 or excellent. Yet teachers' answers to more open-ended questions revealed significant discontentment with the performance of the trainers. When asked what they most disliked about the trainings, over one fifth of teachers expressed displeasure with trainer content knowledge or preparation. When asked what problems the trainers ran into that reduced the trainings' effectiveness, 14% of teachers mentioned trainings starting late, often because trainers arrived late, 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 were unable to answer them, and sometimes the trainers could not solve problems that teachers were asked to solve. Lack of subject experts seemed to be more common in remote locations, where it was more costly and difficult to find and pay for subject experts to participate as trainers.

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,

[34] According to Stallings classroom observation data collected at baseline (involving 10 snapshots during a full class period), of the time teachers spend teaching, on average, 86% of their time is spent on traditional teaching activities (such as explanation/lecture, reading aloud, having students copy information, or running drills), verbal instruction or classroom management, while only 14% is spent engaged in "question and answer or discussion" with students. They were recorded as using blackboards or whiteboards in more than half of the snapshots (57%) while students were involved in group work in only 0.25% of the snapshots.
[35] Of the 98 teachers interviewed over the phone, 20 (20.4%) brought up inadequately trained trainers when asked what they disliked about the training; 11 respondents (11.2%) cited it when asked about the problems that hindered training from being effective; 38 respondents (38.7%) said that trainers should be skilled or better content experts when asked about the ways that can make the training more valuable.

classroom methods or information technology tools. When asked about problems that prevented trainings from being more effective, 15% of teachers mentioned inadequate materials and 11% mentioned inadequate facilities, including a need for labs and information technology equipment. Several trainers also mentioned lack of materials or poor facilities (including lack of electricity) as problems for the trainings. When asked for suggestions of how to improve trainings, 18% of teachers mentioned the need for more teaching materials and 35% mentioned the need for a lab at the training center. Several trainers also suggested that trainings could be improved by guaranteeing adequate facilities and materials. (Also, 9% of teachers mentioned the need for better accommodation.) Given the SSDP emphasis on equipping teachers to use materials and demonstrations when teaching, the lack of materials and lab facilities is especially significant.

Overall, teachers were reasonably positive about the training experience, though their responses suggested much room for improvement. Over 80% of interviewed teachers responded that they found at least half the content to be very valuable, while 28% found at least three-quarters of the material to be very valuable. More science teachers (40%) than math teachers (17%) found at least three-quarters of the content to be very valuable.

Teachers. For the training to succeed, teachers must attend trainings and learn new concepts and practices while at the ETCs. They must also take time to think about how to apply new ideas in their own classrooms. They may have to prepare visual aids, write lesson plans, or think about how to express ideas to their students; this takes time and effort. Then they must follow through and implement new classroom practices well. For teachers to respond in this way, the logistics of training attendance must not entail large difficulties; teachers must come to training with adequate knowledge and skill, ready to understand the training content; the training content must be valuable, and must be presented effectively; and teachers must have adequate time to prepare new lessons, as well as adequate materials and local support. They also require adequate internal motivation or external accountability so that they perceive benefits that outweigh the costs of attending the trainings and implementing new methods.

As indicated in Section 4.1, rates of SSDP teacher training participation among teachers in treated schools at endline were disappointingly low, in part because some of those teachers had joined the schools after the training occurred and in part because invited teachers in the schools at the time of the training did not participate.

Invited teachers may not have attended because they opted not to or because their head teachers did not allow them to participate. Comments during monitoring calls with ETC personnel revealed that some schools did not send their secondary math or science teachers because the trainings were during regular school days, and the schools had no adequate substitutes to teach secondary math or science classes. Our baseline data show that 78% of schools have only one secondary math teacher and 83% have only one secondary science teacher; this suggests difficulty in finding someone with subject expertise to cover for absent teachers. Our data confirm that many schools have difficulty finding substitutes for secondary math or science teachers; over one fifth reported that if a teacher leaves for training, their classes are either cancelled or have no adult supervision.

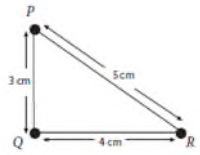
We suspect that some teachers opted not to attend, even though their head teachers allowed and even encouraged them to attend. Some teachers may have thought that the training would be very similar to previous trainings. We cannot quantify this possibility directly, but several training participants reported arriving with little expectation of learning anything new (because of past training experiences and were pleasantly surprised that they did learn useful skills. Others declined the invitation for personal reasons, such as the need to prepare for teacher examinations (an unusual opportunity by which temporary teachers could gain permanent teacher status), travel, illness and family obligations. A few said that the per diems or accommodations were inadequate, suggesting that the costs of participating may have dissuaded some teachers.

Among teachers who attended the SSDP trainings, some may have been ill prepared to benefit from them. Endline data suggest that many may have lacked adequate subject knowledge to fully benefit from the training, which focused on relatively advanced math and science concepts. As indicated in Section 3.4, we indirectly tested teacher subject knowledge by asking teachers to fill out forms evaluating items from the student assessments. In retrospect, we believe the evaluation was not sufficiently challenging, but the results still offer important insights. For each item, the teachers were asked to select the response they believed the question's designer intended as the correct response. Teachers' responses were submitted anonymously, and teachers were (surprisingly) willing to complete the evaluations. Only 2 of the 429 teachers approached refused to complete them. Enumerators reported that teachers in 80% of the schools were enthusiastic about completing the forms, while in the other schools they had no objections to completing them.

There is both good news and bad news in the teacher evaluation results. Encouragingly, many math teachers received perfect or nearly perfect scores. Of the 12 questions included in the math evaluation, we eliminated one from the analysis because it proved to be poorly designed. Thus, the maximum score teachers could obtain was 11 points. Table 14 presents the 11 questions, together with the percentages of responding teachers who selected the correct answer. Over one third (36%) of responding teachers answered all 11 questions correctly, while 72% answered 10 or 11 correctly. Weaker performance on questions 8 and 9 (see Table 14), however, raises questions about the preparedness of a significant minority of math teachers to benefit from the SSDP training. One fifth of teachers gave incorrect answers to question 9, which was a straight-forward algebra problem. Given that the SSDP training was supposed to deal with more advanced algebra techniques, such as factoring polynomials, teachers weak in basic algebra might find it 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.

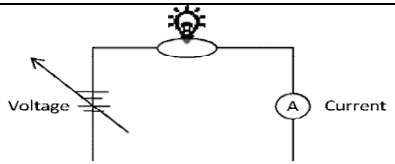
The data raise at least as much concern about science teachers' subject knowledge. Again, many teachers performed well. After removing the question that the fewest teachers answered correctly (which many found unclear and inappropriate to the curriculum), the maximum score was again 11 points. Table 15 lists the 11 questions, together with teacher performance results. Lower total scores for science relative to math may reflect that our science evaluation was more difficult

Table 14: Math teacher responses to evaluations of student assessment items[a]

Assessment item	Number of teachers responding	Percent selecting correct answer[b]	Percent reporting they found item completely clear	Percent reporting they found item very appropriate to secondary curriculum
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	246	97.2	95.1	38.2
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 b. The shape has a right angle c. The shape has equal sides d. The shape has equal angles	246	92.7	67.1	48.4
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	246	97.6	93.0	62.2
4. Which of the numbers below is equal to $\frac{7}{10}$? a. 70 b. 7 c. 0.7 d. 0.07	246	91.5	93.0	57.7
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*	246	96.3	77.3	65.9
6. What is 15 × 9? a. 100 b. 135 c. 130 d. 531	246	91.9	93.9	40.2
7. Which of these is the reason that triangle *PQR* is a right-angle triangle?  a. 3 + 4 = 5 c. 3 + 4 = 12 – 5 b. 5 < 3 + 4 d. 3 > 5 − 4	246	98.8	91.4	84.1
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	246	60.6	80.2	68.7
9. If x + 3y = 11 and 2x + 3y = 13, then y = ? a. 3 b. 2 c. -2 d. -3	246	80.1	90.6	78.0
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	246	91.5	96.7	75.6
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	246	83.7	75.1	64.2

[a] 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 15: Science teacher responses to evaluations of student assessment items[a]

Assessment Item	Number of teachers responding	Percent selecting correct answer[b]	Percent reporting they found item completely clear	Percent reporting they found item very appropriate to secondary curriculum
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	234	93.2	80.3	49.6
2. What is the chemical formula of water? a. H_20 b. NaCl c. NaOH d. H_2O_2	234	97.4	95.7	70.1
3. What is the rate of change in velocity per unit of time? a. Acceleration b. Relative velocity c. Speed d. Velocity	234	92.7	94.4	76.9
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	234	79.1	64.1	64.1
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	234	85.0	70.1	67.5
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	234	34.6	29.1	34.6
7. Which of the following grows from a seed? a. Ant b. Grass c. Mosquito d. Caterpillar	234	88.0	79.5	55.6
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	233	52.4	61.5	55.1
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	234	75.6	62.8	62.0
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	234	59.8	52.1	54.7
11. The symbol of the element nitrogen is: a. N b. He c. O d. H	234	96.2	95.3	71.4

[a] 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?"

than our math evaluation. Only 9% of science teachers answered all 11 questions correctly, but 33% obtained scores of 10 or 11, and 81% obtained scores of 9, 10 or 11. Many science teachers, however, answered incorrectly questions on concepts that seem foundational for understanding the more advanced concepts addressed during the SSDP trainings. One fifth (20%) gave incorrect answers to question 4, which asks about the main function of red blood cells, and which is likely to be an important building block for the SSDP's focus on the circulatory system. Nearly half (48%) incorrectly answered question 8, which relates to evaporation. Again, this is likely an important foundation for the SSDP training's discussion of the more complicated topic of climate change.[36]

These weaknesses in teacher subject knowledge shed both positive and negative light on the SSDP trainings. On one hand, the evidence supports the choice by SSDP training designers to spend time on improving teachers' subject knowledge. Yet on the other hand it raises questions about focusing SSDP trainings on relatively advanced concepts in the 9th and 10th grade curricula, without offering opportunities to strengthen teachers' knowledge of pre-requisite material. Aiming the trainings "too high" may have limited their impact for some teachers. Lack of subject expertise among the trainers, mentioned above, compound this concern. Unfortunately, as indicated above, we found no impact of the SSDP trainings on these measures of teacher subject knowledge.

Even if the SSDP trainings did not improve teacher subject knowledge, they may have improved student learning by improving teachers' pedagogical methods. We learned at baseline that teachers mostly lecture from the blackboard in a traditional teacher-centered approach, with little student engagement and probably only weak knowledge of students' levels of comprehension. Given the focus of the SSDP training curriculum, we might especially hope to see increased use of demonstration methods. Given the nature of the required self-study project work, we may hope to see greater use of lesson plans or of efforts to incorporate local information into classroom and homework activities. Given the exposure of some teachers to the use of group work as a teaching method during the SSDP trainings, we may also hope for some impact there as well. The trainings may also have influenced teaching methods by increasing teachers' enthusiasm for their teaching or by exposing them to the ideas of other teachers. Unfortunately, as we saw above, we detected little or no impact of the trainings on teacher practices (though it remains possible that the program had impacts we were unable to detect, as a result of the imprecision of estimation).

What might explain this lack of the SSDP training impact on teaching practices? Despite some weaknesses in the roll-out of the ETC training sessions, qualitative data suggest that teachers did learn some demonstration techniques that they found potentially useful. 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) responded by referring to using materials, demonstrations or experimental methods. A small percentage (13 %) also mentioned intending to use group work more.

[36] The evaluation forms were submitted anonymously, so we cannot regress scores on teacher characteristics. Yet the correct response rates were very similar when we limited attention to schools where all secondary teachers in the given subject had at least bachelors' degrees in math or science or were permanent teachers. The weaknesses thus are found even for teachers with more education and permanent contracts.

Qualitative evidence suggests that diverse features of the physical and institutional environment may help explain why we find little evidence of change in teaching practices, despite evidence that teachers were exposed to new methods that they viewed as useful. 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 mentioned lack of teaching materials. This suggests that teachers and schools have difficulty obtaining funds for even the relatively low-cost teaching materials made from local resources that were emphasized in the trainings. The great majority of teachers (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 low student attendance (5%), too many students or difficulty managing students (15%), unmotivated students (7%), and lack of teacher time (17%).

Answers to open-ended questions about the video assignment revealed that these demonstration methods can be difficult to implement during class sessions that are only 40 to 45 minutes long, which is the case in all but one of our sample schools. It may also be difficult for teachers to use these more time-consuming teaching methods, given the pressure they feel to complete an ambitious curriculum during the school year. Total instruction time per year varies greatly across schools, primarily because of variation in the number of days school is held (rather than variation in hours of instruction during a typical school week without holidays). The number of instruction days (including any day on which at least one class period of instruction took place) ranges from 184 at the 25th percentile to 208 at the 75th percentile for 9th grade, and from 175 at the 25th percentile to 198 at the 75th percentile for grade 10. In both grades, schools at the 75th percentile had 13% more days of instruction in the last academic year than schools at the 25th percentile. Implementing new methods may have been more difficult in schools with shorter school years.

We also suspect that teachers felt only weak motivation to invest time in preparing to adopt the new methods. Limiting attention to endline teachers who were reported as having attended SSDP trainings both by their head teachers and in the monitoring data, head teachers said they believed that only 34% were highly motivated to try new teaching materials or methods after the training. When teachers were asked during phone interviews if they were either "requested" or "required" to provide reports on what happened during the trainings (after returning to their schools at the end of the training), most (85.7%) reported that they were requested to provide reports, but none indicated that they were required to report. The largest fraction (69.4%) responded that their head teacher had asked for a report, and many (under 44%) were asked by other teachers. Yet only a few (11.2%) were asked by anyone to report what they were doing differently in their teaching as a result of the training. The baseline data indicate that the actors most likely to hold teachers accountable for change are head teachers, but many head teachers are very busy teaching and have little opportunity to observe what teachers do in classrooms. We also know from the video assignment (see Online Appendix B) that at least some teachers welcomed the monitoring and feedback embodied in the video assignment and thought this should be more widespread. When asked about how to improve the trainings, five teachers volunteered interest in having monitoring and follow-up after the training. Many also had positive reactions to the video assignment and expressed interest in getting immediate feedback on their teaching from the focal persons.

Teachers seemed to lack accountability even for the narrower, short-term task of completing the self-study project work component of the SSDP teacher training. Focusing again on teachers that attended the SSDP training according to both head teachers and the monitoring data, 37% were reported by head teachers as not completing 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 one of these projects they had completed, very few could even name any of options! Of the 38 math teachers who attended SSDP math trainings according to both self-reports and monitoring data, only one could describe a project that plausibly may be one of the project options in the official curriculum document. Over one quarter (26%) of these teachers said they did not do these projects and another 58% said they did not know or could not recall the projects they did. Of the 35 science teachers that attended SSDP science trainings according to both self-report and monitoring data, only five could describe projects that plausibly might be one of the options in the official curriculum document. One quarter (25%) of these teachers said they did not do these projects and another 38% said they did not know or could not recall the projects they did, while others gave short phrases that were irrelevant to any of the official project options. Since the unusual projects should have been memorable, we conclude that it is highly unlikely that many teachers gave serious consideration to doing any of the self-study project work. One in-person interview gave further reason to believe that there was little accountability for doing the self-study project work well. The teacher reported: "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.

Trainers' opinions also point to reasons why teachers might fail to adopt new methods. When asked their hopes of how teachers would change their teaching practices after training, most trainers said the use of 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 responded "very likely", while just over half responded "somewhat likely" and 43% responded "somewhat unlikely." When asked about main obstacles to adoption, 61% of the trainers mentioned inadequate motivation, poor attitudes and/or lack of monitoring. Some mentioned especially that teachers are not motivated to introduce new teaching materials because they teach many classes and do not have the time required to prepare and implement new methods. Over one third also mentioned lack of materials or equipment in the schools.

Students. For teaching training to increase students' learning, they must be exposed to improved teaching; be sufficiently nourished to profit from good teaching; have suitable pre-requisite subject knowledge from earlier grades; and be sufficiently motivated to attend class and invest in learning.

While we find little evidence of changes in teaching practices, teachers may have improved their teaching in subtle ways not captured by our measures. Even if students did not receive better teaching, it is useful to check if they were well positioned to benefit from any changes in teaching.

While our data offer only a limited view of students' circumstances, they suggest that inadequate student nutrition is unlikely to explain the program's lack of impact. Most of the students (95%) report having had a meal before coming to school. Family assets (reported above) also suggest

that these families are not from the poorest in Nepal, probably because families that succeed in enrolling their children in secondary level are, on average, better off than the general population.

Our data do, however, 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 students starting the school year with below-grade-level knowledge was a major challenge to teaching and learning in grades 9 and 10. Thus at endline we added assessment items to test students' grasp of material they should have learned in earlier grades and which is needed to learn the more advanced concepts in the SSDP curriculum.

Tables 16 and 17 illustrate some results which suggest that many students have trouble with pre-requisite concepts. For example, consider Question 3 in Table 16, 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 failed to answer this correctly, suggesting that they had not fully learned or retained what they should have in grade 8. More worrying yet, 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 17 raises even more concerns about students' preparation for grade 9 and 10 science. For three of the eight questions, only half or fewer of 9th and 10th graders answered correctly.

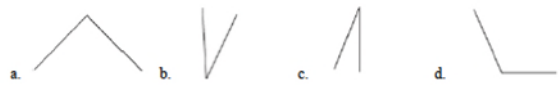
To address concerns that correct responses were low on some questions because students did not take the assessments seriously (rather than because they did not know the material), we used enumerator team reports to identify schools where: (a) the head teacher or teacher encouraged students multiple times to put effort into the assessments; (b) enumerators rated the students as taking the assessments at least moderately seriously; and (c) enumerators reported that there was little distracting behavior during the assessments.[37] When we limit to these schools, we obtain very similar results. For the math questions in Table 16, none of the percentages correct in grade 9 rise by more than 1.5 percentage points, while four of the percentages fall. None of the percentages correct in grade 10 rise by more than 3.5 percentage points, and four of them fall. For the science questions in Table 17, the changes are similar, with only one rising by 3.2 percentage points, no others rising more than 2.1 percentage points, and 10 of 16 percentages fall.

4.4 Promising directions for improving program impact

In the past, the designers and administrators of Nepal's teacher training policies were decision-makers at NCED, Ministry of Education (now Ministry of Education, Science and Technology) and the Department of Education in Kathmandu. Moving forward in the wake of recent government reforms, they will increasingly include decision-makers at provincial or local levels, perhaps with

[37] The enumerators gave separate reports for each section within each grade. For the calculations described here, we selected only schools for which enumerators gave the indicated high rankings to all sections of the given grade within the school. At the 9th grade level, this stringent criterion limits the sample to 80 schools, while at the 10th grade level, it limits the sample to 75 schools.

Table 16: Student performance on below grade-level math questions

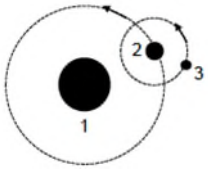
Question	Grade level	9th Graders		10th Graders	
		Number Tested	Percent Correct	Number Tested	Percent Correct
1. If A={a, e, i, o, u}, what is the value of n(A)? a. 4 c. 3 b. 5 d. 2	8 (BL)	6801	74.7	5833	81.7
2. If a square has a side of 6 cm, what is its area? a. 24 cm^2 c. 12 cm^2 b. 36 cm^2 d. 64 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 c. 6 cm^3 b. 4 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 c. 130 b. 135 d. 531	7 (YL)	3397	86.8	2934	91.1
6. What is 6 ÷ 3? a. 18 c. 3 b. 2 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 c. 2 x 5 b. 5 - 2 d. 2 + 2	4 (SLS)	6801	66.9	5833	71.5

Notes: 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 www.timssandpirls.bc.edu). SLS is an assessment item from the Student Learning Study conducted in India (see www.ei-india.com/study_on_student_learning). The percent correct are unweighted averages. 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.

technical assistance from Kathmandu. The evidence presented thus far suggests that, regardless of their level or branch of government, they will need to exercise great care in re-designing and developing a more effective governance structure for new waves of teacher training, so that the funds spent on training lead to significant improvements in student learning.

Teacher training programs achieve high impact on student learning only if their designers and administrators create a curriculum and approach, and a related institutional structure, that satisfy several stringent criteria. First, the training curriculum must focus on material that: (a) addresses true gaps in teacher knowledge or practice that significantly inhibit student learning; (b) equips teachers with new teaching practices that can be implemented without too much additional time and cost; and (c) addresses subject content that students are adequately prepared to understand. Second, the trainings must be rolled out within an institutional structure that equips and motivates

Table 17: Student performance on below grade level-science questions

Question	Grade level	9th Graders		10th Graders	
		Number Tested	Percent Correct	Number Tested	Percent Correct
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: 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 https://www.ei-india.com/Quality_education_study_qes). The percent correct are unweighted averages. 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.

trainers to deliver the curriculum well and fulfill well any other roles they have been given (e.g. in following up with teachers after the training). Third, the trainings must be rolled out through an institutional structure in the schools that equips and motivates teachers to complete any self-study project work and, over the long term, implement improved teaching methods in their classrooms.

With these criteria in mind, we see several strengths in the design and roll-out of the SSDP trainings. First, more than in the past, the SSDP trainings focused on practical, specific teaching

methods rather than on general theories of teaching and learning. In their review of six recent academic reviews of what works to improve student learning, Evans and Popova (2016) conclude that teacher training has the potential to improve student learning significantly, but that merely "providing teachers with general guidance tends not to improve student learning" (p. 260). Rather, teacher trainings are most successful when they are part of larger efforts to improve pedagogical methods used to teach specific subject content. Snilsveit and colleagues (2015) also find that the education programs with the largest and most consistent positive effects on student learning are "structured pedagogy" interventions that provide a package of teacher training and materials for teachers and students, all aimed to help teachers apply specific new methods to teach specific subject content. The teachers in our study also liked the practical focus; when asked what they most appreciated about the SSDP training, half the teachers in our phone interview study provided answers such as "practical way to teach," "use of teaching materials" or "experimental methods."

Another strength of the SSDP secondary math and science trainings was to move teachers away from lecturing at the blackboard, toward a more engaging way to teach, using more demonstrations and experiments. Many programs to improve teaching and learning (in developing and developed countries) encourage teachers to use "active learning" pedagogy. These approaches invite more student engagement – and challenge students to think more analytically – than does traditional lecturing at the blackboard. In their review of pedagogy, curriculum and teacher education in developing countries, Westbrook and colleagues (2013) identify six very effective teaching practices, three of which support the SSDP focus on demonstrations and using teaching materials made from local resources: "frequent and relevant use of learning materials beyond the textbook; open and closed questioning, expanding responses, encouraging student questioning; [and] demonstration and explanation, drawing on strong pedagogical content knowledge."[38]

Other potential strengths are the use of group work during the ETC training sessions, the creation of opportunities for teachers to share experiences with each other, and the required self-study project work. Use of group work in trainings may encourage teachers to use group work in their classrooms, encouraging student engagement. Providing teachers sharing opportunities may allow them to learn from each other's experiences to address practical problems. The self-study project work was presumably to help teachers overcome the hurdle of preparing and implementing new teaching plans. This hurdle seems large, so it is good that the program acknowledges this, even though accountability for completion of this project work seems to have been lacking.

We also see four critical weaknesses in the SSDP design and administration. First, the trainings were rolled out with little attention to equip trainers and hold them accountable for delivery of high-quality training. Trainers appear to need more detailed curriculum guidelines, training of trainers, and more time to prepare the final training content. ETCs also require adequate funding to hire subject content experts and buy teaching supplies to use during training. Our overall assessment, however, is that the teacher training system needs a more fundamental reform, one that equips, supports and motivates ETC personnel and trainers to train more effectively. With ETCs recently being placed under provinces' oversight, the new provincial governments should be encouraged

[38] The other three were: "flexible use of whole-class, group and pair work where students discuss a shared task; use of local language and code switching; and planning and varying lesson sequences." (p. 2).

to consider several changes. First, they should invite ETC personnel, relevant roster trainers (and subject content experts), and teachers to be more involved in developing new waves of training. This may give planners more practical knowledge of the typical barriers to effective training, and may also raise trainers' motivation by giving them more ownership of the trainings. Second, they should conduct more direct oversight over ETCs, trainers and training sessions, including more in-person observation of trainings. It may require a new protocol for following up training sessions, with phone calls (from provincial or local government officials) to the teachers who attended the training, asking about the content covered, instruction quality, likes and dislikes, and any problems that reduced the trainings' usefulness. We raise this possibility because our phone interviews for process evaluation were relatively inexpensive and were useful to learn the *de facto* strengths and weaknesses of the ETC trainings. For phone calls to increase trainers' accountability, it is important to announce a policy of conducting such follow-up phone interviews on a regular basis. Still, efforts to improve training implementation should be done carefully, since efforts to improve governance in other contexts have failed or even backfired (Finan, Olken and Pande 2015).

Second, the trainings provided teachers with inadequate motivation and mentoring to implement new methods in their classrooms after the ETC training sessions. While further experimentation can be useful to identify effective reforms, we believe that NCED and provincial policy makers should consider new training modes that emphasize mentoring of teachers in their schools. Evans and Popova (2016, pp.260-261) find that "one-time in-service trainings at a central location … are not among the [teacher training interventions] found to be highly effective." By contrast, Conn (2017) finds evidence that "pedagogical interventions involving long-term teacher mentoring or in-school teacher coaching … produce a sizeable (albeit not always significant) effect on student learning…". Similarly, Albornoz and colleagues (2019) compared different trainings of science teachers in Argentina. Teacher training combined with on-going coaching had a much larger impact on students' science test scores than did training alone. For SSDP training policies, this may consist of follow-up visits to observe teachers implementing new methods in their classrooms after ETC training sessions, with explicit feedback, discussion of challenges, and possibly multiple visits over weeks or months. Other modes of mentoring can also be explored, such as using social media to connect secondary math or science teachers within districts, and developing internet-mediated mentoring sessions to teachers from centralized experts in subject content or pedagogy.

Third, SSDP teacher training was not delivered as part of a larger package (such as distribution of lesson plans and teaching materials) that teachers could easily adopt in their classrooms. To implement the demonstration methods promoted in the SSDP trainings, teachers would have to spend much time preparing new lesson plans and creating new teaching materials. Lack of time, of motivation for preparation, and of funds to acquire even very modest local resources, seem to have hampered adoption of new teaching methods. One rigorous study finds that teacher training combined with new materials and lesson plans raises student learning more than teacher training alone. The Piper and colleagues (2018) compared three interventions in Kenya: a teacher training program, the training program plus revised student textbooks, and a program combining training, revised textbooks, and structured lesson plans. The program combining all three increased student learning the most. The value of combining teacher training with lessons plans and materials is suggested by the success of "structured pedagogy" programs (Snilsveit *et al.* 2015).

Finally, we strongly suspect that the trainings focused too exclusively on the teaching of relatively advanced grade 9 and 10 subject content. While our evidence suggests that some teachers would benefit from training to improve their mastery of this content, exclusive focus on this may render the trainings unhelpful for other teachers, who need remedial work with lower-grade subject content before progressing to grade 9 and 10 content, and for the many students entering grades 9 and 10 with large deficits in understanding lower grade level math and science concepts. We suggest encouraging "teaching at the right level," both by teachers in their classrooms and by trainers when teaching teachers. Interventions that group students at different levels of current performance and provide differentiated instruction to those groups are promising. Duflo, Dupas and Kremer (2011) showed that tracking students into "high" and "low" achievement increased math and literacy scores of both groups in Kenya. A similar intervention is to provide remedial instruction to students who are falling behind; Banerjee and colleagues (2007) showed that, in India, providing low-performing Grade 3 and 4 students with two hours of daily remedial instruction greatly increased their reading and math scores. Evans and Popova (2016) point out that training teachers to use formative assessment and targeted instruction within classrooms raises student learning. Such studies show the value of carefully targeting instruction to student learning levels, but are not directly applicable examples as they focus on primary instead of secondary education.

The challenges our government collaborators faced in rolling out the SSDP trainings motivate three smaller suggestions. First, some teachers invited to the SSDP trainings did not attend because their schools had difficulties finding substitutes to teach their classes while they were at the trainings for two weeks; in some cases, teachers did attend but their students were left without supervision. This problem is especially acute for secondary math and science teachers, who are in short supply. This implies that one should consider scheduling trainings during school vacations or using new training modes that reduce the need for teachers to be away from their classes. Second, some teachers may have chosen not to attend because they expected not to learn anything at the training. More effort may be needed to "market" trainings to teachers, convincing them that there will be new material, and conditions will be comfortable and conducive to learning. Third, some teachers reported difficulties in implementing the new teaching methods promoted at the SSDP trainings because they require class sessions longer than the 40- to 45-minute sessions that are typical in Nepal's secondary schools. Schools should consider whether altering class schedules to allow teachers of some subjects to have some longer class sessions, may be useful.

In this section we identified strengths and weaknesses in the design and implementation of the SSDP teacher trainings, and suggested ways for policymakers to improve the design and implementation of future trainings. We should point out, however, that none of these suggested changes is guaranteed to succeed. We think that policymakers should start experimenting with innovative training programs on a small scale, in conjunction with careful evaluation and iterated improvements, developing the most promising program designs before taking them to scale.

5. Cost analysis

Estimating costs for the SSDP teacher training intervention is difficult, because it was rolled out by government institutions that use shared resources for many interventions and made use of

personnel and institutional investments from previous waves of teacher training. It is therefore important to emphasize that our cost estimates are rough approximations.

We estimated the SSDP training costs in three steps. (Supplementary Table 3 in Appendix F summarizes the steps.) First, we listed all the activities required to develop the training curriculum and roll it out at the government institutions in place before the training began. This list has three categories: (1) costs incurred only once for the entire program, regardless of the number of schools in which the program is rolled out; (2) costs incurred only once per ETC, regardless of the number of schools and teachers that participated in the program in the ETC catchment area (usually one to three districts); and (3) costs incurred each time the ETC rolls out a training session for about 20 teachers. Our second step was to estimate the Category 3 costs from actual training session costs in one of the seven ETCs that continued operating after the recent reform. Because the Finance Ministry and Ministry of Education, Science and Technology tend to provide nearly identical budgets to all ETCs implementing the same training sessions, we believe it is reasonable to extrapolate from this ETC to the others. We recorded both Category 2 items as zero, because there were no new activities of this type as part of the SSDP trainings, but we include these line items in the table, since such costs may occur if similar trainings are rolled out in other contexts. The third and final step, which yields numbers with the most uncertainty, was to estimate Category 1 costs at the national level and some Category 3 administrative costs at the training session level, working with one or our local partners, the Center for Policy Research and Consultancy, which has long experience collaborating with, and performing research on, education institutions in Nepal. With this partner we developed rough estimates of the time required to complete each task and the pay grade at which that task would be done, using government pay scales. Note that international organizations that work with education policymakers in Nepal, such as the Asian Development Bank, United Nations Children's Fund and the World Bank, also contributed to the design of the SSDP trainings, and we have not included the costs of their contributions.

We estimate that the total cost for all 16 study districts to roll out two trainings (one in math, one in science) in the 14 ETCs relevant to the study was about $73,000 (using a NR. 110 per dollar exchange rate). Each training had about 20 teachers. These costs work out to about $2,600 per training session, so $130 per teacher. The average number of grade 9 (grade 10) students per section (and so, usually, per science or math teacher) is 50 (43). The cost per student of training their teachers in one subject is thus about $2.60 ($3.00) per grade 9 (grade 10) student.

According to our calculations, by far the largest cost of the trainings is the per diems and lodging for the participating teachers. This suggests that adding more trainings in these 14 ETCs for additional sets of 20 teachers would entail marginal costs nearly equal to the total cost per teacher. Expanding the trainings to catchment areas of new ETCs would likely cost more, especially if those ETCs conducted trainings of trainers and made other one-time preparations.

To put these costs into perspective, we turn to Damon and colleagues (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) range from $19 (excluding administration costs) for a girls' scholarship program, $24 (excluding administration

costs) for private school vouchers, $15 for a relatively expensive computer-based program, $3 to $9 (excluding administration costs) for student incentive programs, to $2-3 (excluding administration costs) for teacher incentive programs. Overall, while the SSDP intervention has a relatively low per-student cost, it is a cost per student that, 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 by an intervention that can increase student learning.

6. Discussion

This mixed methods evaluation estimates the impacts of the SSDP training for grade 9 and 10 math and science teachers on student learning and evaluates the strengths and weaknesses of the program's design and implementation. Based on a randomized control trial involving 203 schools in 16 districts from all provinces of Nepal, we find no evidence of positive program impacts on learning. In fact, our estimates rule out anything more than a small positive impact, and even suggest negative impacts, especially for better students (as measured at baselined). Using both qualitative and quantitative data, we described key strengths and weaknesses of the program's design and implementation (Section 4.3). We summarize the findings and recommendations in Section 7 below. This section discusses the study's limitations and external validity, describes our efforts to encourage evidence uptake, and draws lessons for future evaluations of education interventions (including but not limited to teacher training interventions) in Nepal.

6.1 Limitations and external validity

Limitations. Challenges during the intervention's roll-out led to two main limitations of the study. First, the delayed roll-out reduced the time between roll-out and endline data collection, reducing the time over which any improved teaching could increase learning. The average time between the math training in the treatment schools and the endline testing was 11 months, with a range from 9 to 14 months. The average time between the science training and the endline testing was 9.4 months, with a range from 7 to 15 months. While we believe that this was enough time for most teachers to implement improved teaching methods across most of the curriculum content, and so enough time to affect student learning, perhaps with additional time teachers would have integrated more new methods into lesson plans, and gained skills by practicing the new methods, leading to greater impacts.[39] Extending the study by another year would have allowed us to estimate impacts after two years for the students in grade 8 at baseline. Yet this was not feasible, since policymakers were reluctant to withhold training from control schools for another year.

[39] This variation in months of exposure to trained teachers suggests that it may be of interest to replace the "Treat" dummy variable in the regressions in Table 5 (and other tables) with the number of months or days that students were "exposed" to trained teachers. This would capture the variation in students' exposure to the treatment that is missed by the "Treat" dummy variable, and a quadratic specification could be used to check for heterogeneity in the impact by length of exposure. Yet this variation in days of exposure was not randomly assigned, so it is unclear whether a causal interpretation can be given to the results. Even so, Online Appendix Table A.12 provides such estimates for Table 5 (full sample, using the full assessments). The linear estimates shown (quadratic terms were all far from significant) are very similar to those in Table 5, with negative and at most marginally significant impacts for grade 9 math and science, and even smaller and completely insignificant impacts for 10th grade math and science.

Second, teacher turnover, and especially teachers' non-take-up of training, led to low rates of completed training among teachers in our study schools. This reduction in numbers of teachers trained in treated schools reduced the precision of our estimates of treatment impacts.

External validity. Compared to many studies, this evaluation was well designed for external validity, at least within the policy environment that prevailed during the evaluation's design phase. We collected a (nearly) nationally representative sample of public schools, involving 16 districts from all areas of Nepal.[40] In addition, this intervention was rolled out by the government, largely through the institutions and procedures in place for government education policy prior to the study.

Some departures from standard practice were needed, however, in order to carry out high quality, informative research under a budget constraint. Some teachers invited for training in our study would have been excluded under status quo ante protocols, which prioritized including teachers with permanent contracts who had not been trained under a previous education policy (the SSRP). Also, schools received explicit invitations to send their teachers to the trainings, while the status quo ante protocol was to wait for teachers and schools to request training. More importantly, as the NCED did not collect systematic data from ETCs on training dates, training attendance and other monitoring indicators, and because the NCED seemed not to have communication channels or personnel in place to ensure that the ETCs invited teachers from treated schools and not from control schools, research team members made frequent phone calls to the ETCs for data gathering and oversight. This may have raised the ETC personnel's perceptions of accountability somewhat relative to status quo procedures, though our impression is that this effect was small.

A larger external validity question concerns Nepal's recent dramatic government reform (which accelerated after our intervention rolled out); it created new local and provincial governments and seeks to shift governing authority from the federal (central) level to the provincial and especially local levels. (Schaffner, Glewwe and Sharma 2019b, summarize the implications for education policy.) In principle, the new structure shifts the responsibility for providing basic and secondary education to local governments, with federal institutions such as the former NCED playing only a facilitating role. In fact, the exact form of the new institutional structure is still unclear. Legislation to reform the education sector is not yet enacted, and many positions in the new provincial and local governments' education departments are unfilled. While this reform means that trainings identical to those we evaluated will no longer be rolled out, limiting our external validity in one sense, in a broader sense it creates a good opportunity for policymakers at all levels to learn from the study and design new institutions that pay greater attention to issues highlighted in this report. The results are also valuable outside of Nepal, as they suggest possible ways of improving the performance of training-center-based in-service teacher training programs.

[40] In addition to increasing our impact estimates' external validity, use of a nationally representative sample increases the value – for policy makers and education researchers – of our datasets, since this is one of the first efforts in the last 15 years to collect systematic, quantitative data on secondary education in Nepal.

6.2 Policy and programme relevance: evidence uptake and use

At all stages of this project, we strived for a strong partnership with government collaborators. As explained in Section 2, we worked extensively with policymakers to identify the study intervention and develop the evaluation questions, hoping to build buy-in and engagement. The strong support and leadership of Dr. Teertha Dhakal, then the National Planning Commission Joint Secretary, were essential for bringing together representatives from the National Planning Commission; Ministry of Education, Science and Technology; Department of Education and (later) the NCED, not only for participation in large workshops and trainings, but also to participate in a "technical committee" that provided specific input to the evaluation, especially while preparing and executing baseline data collection and rolling out the intervention. Government participation waned after roll-out and through preparation of endline data collection, as accelerating government federalization brought transfers to new positions for many government officials initially involved with the study. Interest revived somewhat during analysis of endline data, with a good turnout at our final dissemination workshop in August of 2019, drawing representatives from the main federal-level government agencies involved in education policy, and planning members from four of Nepal's seven new provincial governments. Policymakers' reactions during this event indicated that they heard the main message (of ineffective trainings and the need for overhaul) and appreciated the need to act on the findings. Yet whether and how they will act is unclear, given uncertainties about the large on-going government reforms. The study may also affect policy through the School Sector Development Program Technical Working Group, which brings together senior government officials and international development partners, and which discussed our preliminary report, and through staff from National Planning Commission and other organizations for whom we provided an impact evaluation workshop in January, 2020. Lessons for collaboration between researchers and government officials on major evaluations, based on our experience, are in Online Appendix C.

6.3 Challenges and lessons

Sections 3.1 and 3.3 described challenges encountered when rolling out the teacher training intervention and examined the possible policy implications (summarized in Section 7 below). In this section we describe eight challenges we faced while implementing the evaluation and discuss the implications for future evaluations of education policies in Nepal.

First, we found it difficult to obtain even the basic data required to construct a sample frame. In principle, the Department of Education collected administrative data on numbers of students and teachers by grade for all schools. In practice, government officials were slow to share these data with us. Other data we needed, such as accurate records on past teacher training experience of current teachers, were unavailable. A well-functioning Education Management Information System could have raised the quality and speed, and reduced the cost, of our evaluation.

Second, after finding problems in the design of the baseline student assessments, we devised a new approach for developing assessments that we believe led to significant improvement. This process, described in Section 3.4, involved a team of international and local assessment experts.

Third, suspecting that teachers may be reluctant to participate in assessments of their subject knowledge, we devised an indirect method that seemed to work well. For a description of the instrument we asked teachers to complete anonymously, see Section 3.4.

Fourth, having difficulties obtaining detailed descriptions of the curricula and methods used for the SSDP trainings, we designed and implemented a telephone interview study, involving some trainers and most of the teachers who had been interviewed at baseline and had subsequently attended SSDP trainings. This was a relatively low-cost component of our study, but the answers to closed-ended and short-answer open-ended questions were very useful for understanding the training content, methods, and quality of the training sessions. The telephone study was made feasible and inexpensive by requesting teachers' phone numbers during baseline data collection.

Fifth, we found that the combination of constraints imposed on us by government collaborators, our funder, and our Institutional Review Boards tended to create severe timing problems that may have reduced the study's quality. To be evaluable and ready for evaluation funding, interventions must be ready or nearly ready to roll out, but policymakers are reluctant to delay rolling out an intervention that is ready to go. This can result in a short window – between the time that funds are awarded and the deadlines for roll-out set by policymakers – to design samples and develop baseline instruments. These time constraints bind even more severely due to Institutional Review Board processes that cannot start until the sample and instruments are finalized. We were fortunate to find funds from another source to finance a preliminary qualitative study before 3ie approved the full proposal, which helped us meet our tight schedule for baseline development. We suggest that 3ie and other funders consider modifying their funding rules to permit the release of some funding for preliminary qualitative research and baseline development before approving an evaluation's full funding, at least when previous funder interactions with the research team suggest a high probability that a full proposal will be approved.

Sixth, as indicated above, we believe we could have improved our study's quality and impact if we had embedded a research team member in a key policy implementing agency for a few days per week during much of the study period. Even if we had realized the importance of this before developing our proposal, we could not have done this due to budget constraints. We suggest that 3ie and other funders set funding limits that accommodate embedding of a research team member in the implementing government agency when funding evaluations of government programs.

Seventh, despite taking what we considered a conservative approach to power calculations (for example, assuming an unusually high value for the intra-cluster correlation coefficient) and thus choosing a relatively large sample size, we believe our sample size was too small to produce adequately precise impact estimates.[41] This leads us to offer one practical suggestion regarding

[41] Our main sample size calculation aimed for an MDE of approximately 0.2 standard deviations on assessments for 9th and 10th grade math and science students, with significance level of 95 percent and power of 80 percent. Based on analysis of earlier nationally representative academic achievement tests, we set the intra-cluster correlation coefficient to 0.65, which is unusually high and implies the need to include many schools in the study. We assumed the inclusion of at least 30 students per school (in any subject and grade). While the MDE in a sample of 200 schools based on these calculations was slightly higher than the target, we anticipated being able to improve power by including baseline student test scores and district-

power calculations and to highlight a potential flaw in current power calculation practice. The practical suggestion is to integrate conservative predictions regarding program up-take into power calculations. While we made conservative assumptions regarding several power calculation parameters, we failed to consider possible lack of compliance among teachers assigned to treatment. We assumed perfect compliance, but in practice 40% of endline math teachers and 58% of endline science teachers did not attend training, increasing imprecision in estimation.

The potential flaw in power calculation best practice that we wish to point out concerns MDE targeting. Current best practice calls for calculating the sample size required to achieve, say, an 80% probability of obtaining a statistically significant impact estimate (at, say, the 95% confidence level) if the true effect is equal to an MDE selected by the researchers. This MDE is often an effect size that is just large enough that policymakers would take it as indicating an effective program (and thus want to be able to detect it). The MDE is, therefore, typically modestly large. Pressure to prove to funders that samples of adequate size are feasible given grant size limits may create incentives to raise these target MDEs even higher. Yet this MDE-based approach to sample size calculations sets a low standard regarding the likely precision of impact estimates. An estimate equal to the MDE can meet the standard of statistical significance even if the confidence interval for that estimate ranges from only just slightly above zero to nearly twice the size of the MDE. When MDEs are modestly large, estimates with confidence intervals this wide are not precise enough to provide useful calculations for benefit and cost comparisons. Further, if the true effect is zero, confidence intervals this wide (ranging from nearly as high as the MDE down to the negative of that effect size) are not narrow enough to rule out modestly large effects, and thus are too wide to allow definitive conclusions regarding a program's ineffectiveness. While in many cases our standard errors were small enough to allow detection of true effects equal to our MDE, and in some cases allowed us to rule out anything more than small positive impacts (because our point estimates were negative), the standard errors were nonetheless larger than desirable for precise impact estimation. Estimates of heterogeneous effects were even more imprecise. This inadequacy raises important questions about funding ceilings and what sorts of programs it is feasible to evaluate. Had we cut our MDE in half, in order to obtain more precise impact estimates, we would not have been able to proceed with the evaluation; it would not have been feasible to include a large enough sample while remaining under the 3ie budget limit for this evaluation.

Eighth, we believe that the current government reform, which aims to shift policy making from the federal level to provincial and local levels, constrains the types of evaluation that will be feasible in Nepal for the next few years. The reform introduces great uncertainty regarding how long policymakers will remain in their current positions, which policies and programs will continue in their current forms, and for how long, and which institutional structures will provide oversight for specific policies. This raises the risk that evaluations will lose institutional support and be ended before reaching conclusions, especially if they require collaboration with many government institutions for long durations. In this environment, we suggest that evaluations be somewhat smaller in geographic scope and less complex than our study. Several more tractable types of evaluation can be valuable. First, since costly large-scale evaluations should be done only for programs with

stratum fixed effects as controls. In practice, across grades and subjects at endline, the intra-cluster correlation coefficient s ranged from 0.063 to 0.095.

high probabilities of success, policy designers need evaluators to help them develop programs a high probability of success. In particular, evaluators could help policy designers raise and answer important questions about proposed programs' objectives, theories of change, and design details, and then help them design and analyze small-scale policy experiments. The process of design modification and experimentation should continue until programs have a high success probability. Second, for programs that are ready for rigorous evaluation, evaluators should consider implementing experiments in a single province (or an even smaller geographic area), to reduce administrative complexity. This will allow evaluators to maintain good communication with relevant policymakers and program personnel. Third, evaluators should look for possible innovations in program governance with the potential to improve the implementation of specific programs, and should consider rigorous study of the impacts of those changes on program implementation outcomes. For example, an evaluation might study the impact of specific changes in the monitoring and supervision of trainers on training session quality outcomes, such as trainer promptness and time use during training sessions, quality of training materials, and teacher satisfaction. We suggest this as we believe that better governance should be a high priority, and we suspect that impact evaluations could be done more easily and quickly when the impacts of main focus are impacts on implementation outcomes rather than final outcomes (such as student test scores).

7. Conclusions and recommendations

This mixed methods evaluation estimated the impacts of Nepal's SSDP trainings for secondary math and science teachers on teacher subject knowledge, teaching practices and student learning, and analysed the strengths and weaknesses of the program's design and implementation. It combined an RCT of 203 schools in 16 districts with several qualitative research components, including the collection of monitoring data, a "small N" study involving in-person interviews, and a "larger N" part qualitative, part quantitative study involving telephone interviews of teachers and trainers who participated in the SSDP trainings.

We found no evidence that the SSDP training for secondary math and science teachers raised student test scores. In fact, our main results allow us to rule out anything more than small positive effects, and in some cases we estimate statistically significant negative impacts. We find weak but suggestive evidence that any negative effects are largest for the students who were highest performing at baseline. At about $130 per teacher, or $2.60 to $3.00 per student, the cost of the SSDP trainings is similar to that of interventions that have been found to raise student learning significantly in other contexts. We thus conclude that Nepal's policymakers must improve teacher trainings or replace them with demonstrably more effective interventions. We hope the findings are useful for education policymakers at the federal level, education officials in the new provincial and local governments, and for other stakeholders in Nepal's education sector, such as local non-governmental organizations and international development partners. The findings also point to potential weaknesses in training program design and implementation that education policymakers in other low- and middle-income countries should look for and address.

Drawing on qualitative and quantitative evidence, we describe five sets of problems that may explain why the SSDP trainings did not improve student learning. First, weak governance likely

reduced the quality of the ETC trainings. It appears that trainers were given inadequate time and guidance to prepare training materials, were given no "training of trainers," and in some cases lacked relevant teaching materials. Some ETC's trainers lacked adequate expertise in math and science. Second, scheduling training sessions on regular school days may have prevented some teachers from participating because substitute teachers were unavailable to teach their classes during the trainings. Teachers' low expectations of the novelty and value of the SSDP trainings may also have lowered participation. Third, we find evidence of serious weaknesses in some teachers' pre-requisite subject knowledge, which may have impeded them from grasping training content focused on advanced math and science concepts. Fourth, few teachers seem to have completed the post-ETC self-study project work or adopted new classroom teaching methods, and our evidence suggests two possible explanations: (1) Teachers' lack of accountability for the time-consuming development of lesson plans and teaching aids; and (2) Teachers' lack of budgets for needed teaching materials. Finally, we find that many students enter grades 9 and 10 with below-grade-level math and science skills. SSDP trainings focused entirely 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.

Our study has two limitations. First, it is possible that SSDP training impacts grow over time, and we estimated impacts after only one year. Second, the training completion rates in our study schools at endline were unusually low, reducing precision, due to high teacher turnover and low teacher take-up of the training invitations.

Compared to many studies, this evaluation was designed relatively well for external validity, because we use a nearly nationally representative sample of schools to study an intervention rolled out through the institutions that were responsible for government training at the start of the study period. A dramatic government reform, however, recently shifted responsibility for basic and secondary education to new local governments, so that trainings identical to those we evaluated will no longer be rolled out, limiting our external validity in a narrow sense. Yet in a broader sense the reform creates valuable opportunity for policymakers at all levels to learn from the evidence and pursue improvements in teacher training program design and implementation.

Considering our evidence on the problems that may have reduced the SSDP training program's impacts, we recommend that policymakers in Nepal, and in other countries where policymakers find evidence of similar weaknesses in training program design and implementation, experiment with changes of the following sorts: (1) Allocating more training time for methods to identify, and differentiate instruction for, students entering grades 9 and 10 with below-grade-level subject knowledge; (2) Re-designing trainings to better accommodate teachers with gaps in pre-requisite subject knowledge; (3) Combining trainings with distribution of related lesson plans and materials (to reduce potential barriers to adoption of new teaching methods); (4) Connecting trainings to periodic classroom visits (either in-person or virtual) by mentors or coaches who can advise, monitor and hold teachers accountable for improved teaching; (5) Improving the way trainers are trained, equipped and motivated to deliver high quality trainings; (6) Scheduling trainings outside of school hours or during school breaks to increase participation; and (6) Increasing efforts to motivate teachers for training by informing them about the novelty and value of the new training.

Appendix A: Field notes

New Era Pvt. Ltd., which has been in operation since 1971, was hired to collect both baseline and endline quantitative data. The collection of the baseline data took place between August 2017 and January 2018. For this data collection, New Era created 18 teams, each consisting of two enumerators and one supervisor, to administer questionnaires and student assessments during visits to schools that usually lasted two to three days. All the enumerators and supervisors had either a Bachelor's or Master's degree, and many had experience in the education sector. Travel between schools usually took between several hours and a full day. The data collection (which consisted of administration of questionnaires and assessments) was done in two phases, with a gap of several weeks in between the two phases, to accommodate the holidays that most schools took for at least three weeks between September and October of 2017. Multiple teams sometimes worked in the same district at the same time to complete the fieldwork in all the schools within the district before the holiday break. The first phase of the fieldwork, which took place in 12 districts, started on August 17, 2017 and was completed on Sept. 18, 2017. The second phase of fieldwork in the remaining four districts -- Baglung, Morang, Nuwakot and Parsa -- started on October 24, 2017, and was concluded on Nov. 27, 2017.

For the baseline data collection, New Era also selected 12 individuals with supervisor-level qualifications to perform Stallings classroom observations, which took place at a minimum of two weeks after the initial school visit. While schools were informed that such visits would take place, they were not told the specific date for the classroom observation visit. The Stallings classroom observations were conducted from August 30, 2017, to January 18, 2018.

All baseline enumerators participated in a 10-day training led by the research team. Those selected to perform classroom observations also received an additional three-day training course from August 10th through 12th, 2017, led by Rashmi Menon and Anuja Venkatachalam, Senior Research Associates at J-PAL South Asia.

For the endline data collection, New Era created 17 teams, with two teams assigned to Morang district (where the sample of schools was twice as large), and one team assigned to each of the other districts. A team's visit to a sample school usually lasted one and a half to two days. Again, each team had a supervisor and two other enumerators, and all the enumerators and supervisors had either a bachelors or a master's degree. Enumerators and supervisors were trained by the research team over 10 days between January 28th and February 9th, 2019. The research team worked closely with New Era to ensure that both enumerators and supervisors fully understood how to administer the questionnaires, and the protocols to follow for introducing themselves, administering informed consent, ethical treatment of human subjects, administering assessments, and completing other forms.

When first arriving in a study district, each baseline and endline enumerator team first visited the offices of the District Education Development and Coordination Unit (which were called the District Education Offices at baseline), with a letter of introduction from the Ministry of Education, Science and Technology to inform officials about the research. The officials there were largely cooperative

and responsive and drafted letters to our sample schools attesting to their support for the study. The field team then took the same letters of introduction from Ministry of Education, Science and Technology and the letter from the District Education Development and Coordination Unit (or District Education Office) to the Head Teacher or Assistant Head Teacher at each sample school.

During the baseline, the following questionnaires were administered:

- A head teacher questionnaire, administered mostly by a supervisor using a tablet.
- A questionnaire for teachers of 9th and 10th grade math and science, administered by either a supervisor or an enumerator using a tablet.
- A student questionnaire, which students in grades 8 and 9 were asked to fill out on their own on paper copies.
- Student assessments in 8th and 9th grade math and science.
- Measures of classroom teaching practices and student engagement derived from use of the Stallings method of classroom observation (*World Bank, 2015*).

During the endline, more instruments were administered. The main questionnaires and assessments were:

- A head teacher questionnaire, administered mostly by a supervisor using a tablet.
- A questionnaire for teachers of 9th and 10th grade math and science, administered by either a supervisor or an enumerator using a tablet.
- A student questionnaire, which students in grades 9 and 10 were asked to fill out on their own on paper copies.
- Student assessments in 9th and 10th grade math and science.
- A School Management Committed respondent questionnaire, administered mostly by a supervisor

In addition, enumerators at endline completed forms related to:

- Teacher attendance from school log books
- Identity and movements of teachers in and out of the school between baseline and endline
- Tracking of baseline students (a form that also included questions for head teachers about students' ethnicity and performance on the previous year's end-of-year exams)
- Teacher Evaluations of Student Assessment items, which at least one teacher of grade 9 or 10 compulsory math and at least one teacher of grade 9 or 10 science were requested to complete anonymously on paper copies

The Head Teacher/Assistant Head Teacher was asked about who he or she could assign to help the field team member with the Teacher Turnover Form and the Teacher Attendance from Log Book Form. Often, they volunteered to provide this help themselves.

Early in the endline field work the survey firm discovered that the numbers of students present at endline in some schools were significantly greater than the numbers present at baseline. Concerned that the teams would run out of the paper copies of questionnaires and assessments that they carried with them to their study districts (some of which are quite remote), the survey

firm, in consultation with the research team, implemented the following rule: *In all schools, administer assessments to all students from the baseline study who are present at endline and who assent to participation. In schools where the number of "new" students (i.e. students not present at baseline but present at endline) was less than or equal to 10 % of the baseline number of students, administer assessments to all new students who assent to participation. In schools where the number of new students exceeded 10 %, administer assessments to only one-third of the "new" students.* In a few cases in which the rule dictated including only one third of the new students, head teachers requested that all students be included, and the enumerator teams complied. In the end, all students who were present in the school at endline were included in the study in 71% of the schools for grade 9 and 83% for grade 10.

The major problem that the survey firm encountered were the unexpected public holidays and internal school exams. These problems were particularly severe during the baseline as the survey team did not have the contact numbers for head teachers or teachers in the schools. The survey team would be made aware of these holidays only after reaching the school. Given the remoteness of many schools in our sample, it made more sense for the teams to wait in the school until the students were done with their exams, or the holiday was over, rather than attempting to visit another school.

During the endline survey, coordinating with schools was easier given the teams' access to phone numbers of head teacher and teachers, but unexpected school closure was still a problem. There were 5 or 6 unexpected school closures that the teams faced during fieldwork (public holiday due to the death of a minister, strike by *rahat* teachers, strike by a political party, and a public holiday to commemorate the anniversary of province government formation). Since these holidays were declared at the last moment, it posed severe logistical challenges to the field teams, not least because they were under severe time pressure as schools were about to be closed for the year. In addition, the field teams had to change their plans in some schools since it coincided with the schools' exams. There were also instances where it was difficult to contact head teachers especially in remote areas via phone due to cellphone network problems and/or outdated phone numbers.

Appendix B: Survey instruments and other evaluation tools (qualitative and quantitative)

Instruments.zip

Appendix C: Pre-analysis plan

PAP_April30_2019_final.zip

Appendix D: Sample Design and Calculation of Population Weights

This appendix describes the selection of our study sample and our construction of school-level population weights, which can be used to produce statistics based on the sample of 203 schools that are representative of the 16 districts chosen for the study. The secondary schools in these 16 districts, in turn, are nearly representative of all secondary schools in Nepal (as discussed in the main text).

I. Selection of 203 Schools

As described in the main text, the 203 schools in the study were chosen using a 2-step process. For complete details, see subsection IV.A in Schaffner, Glewwe and Sharma (2018). First, 20 districts were randomly selected to be representative of 65 of Nepal's 75 districts. These 20 districts were later reduced to 16 districts, as explained below. Second, within the 16 districts, schools were randomly selected from two strata within each district.

Selection of 20, and then 16, Districts. When the sample was drawn in January of 2017, Nepal had 75 districts. (It now has 77.) Of these 75 districts, 10 districts (Bajura, Dolpa, Humla, Kalikot, Manang, Mugu, Mustang, Nawalparasi, Rukum and Taplejung) were dropped from consideration due to extreme remoteness or other challenges that would raise data collection costs. According to a database provided by the Ministry of Education, in 2016 there were 8,681 public schools (which are called "community schools" in official statistics) that include at least grades 1 through 10 in Nepal. These combined primary-secondary schools educate 97% of Nepal's 9th and 10th grade public school students and are considered to be the model of what most schools will soon be in Nepal (see Schaffner, Glewwe and Sharma, 2018). Of these, 497 were in the 10 excluded districts, so the 65 districts that are represented by the sample included 94.3% of Nepal's public schools with at least grades 1 through 10.

These 65 districts were grouped into 14 strata by "terrain" (mountains = 1, hills and plains = 2), and by the 7 provinces in Nepal, each of which contained, on average, about 10 districts. The probability that any given district was selected into the stratified random sample of 20 was proportional to the number of public schools with grades 1 through10 in the district. The one exception to this was that, at the request of policymakers in Nepal, districts in Province 6 were given a "double probability" of being selected (see Schaffner, Glewwe and Sharma 2018 for details). The 20 districts randomly selected were: Solukhumbu, Jumla, Khotang, Panchthar, Morang, Parsa, Mahottari, Nuwakot, Kavrepalanchok (Kavre), Sindhuli, Chitwan, Lamjung, Tanahun, Baglung, Arghakhanchi, Kapilvastu (Kapilbastu), Dailekh, Salyan, Baitadi and Achham.

Constraints on the personnel time available for working with districts to obtain sample frame data necessitated dropping four districts. After assessing data availability and the challenges associated with each district, and wishing to retain at least one district per province, the research team together with policymakers chose to drop Baitadi, Khotang, Mahottari and Tanahun districts. See Schaffner, Glewwe and Sharma, 2018, for further details.

Selection of 203 Schools within the 16 Districts. Power calculations suggested that we aim for a sample of approximately 200 schools, of which half would be randomly assigned to the treatment group and half to a control group. After further discussions with Nepalese officials in the summer of 2017, we decided

to stratify the sample not only by district, but also by schools' "priority" or "non-priority" status, selecting two-thirds of the sample from among priority schools and one-third from non-priority schools. Priority schools are those for which official hardcopy NCED records showed that no permanent math or science teacher (and no math or science teacher whose permanent or temporary status was unknown) had completed SSRP training (i.e. training under the previous seven-year education plan). Non-priority schools are those for which the records showed that at least one permanent secondary math or science teacher (or one math or science teacher whose permanent or temporary status was unknown) had completed SSRP training. To facilitate the two-thirds/one-third stratification, we required a number of schools per district that is divisible by three. It was also decided that, within each district, half of the schools would randomly be assigned to the control group, one fourth would be randomly assigned to the "standard" treatment group, and one fourth randomly assigned to the "treatment plus video assignment" group. This rendered it convenient to select 12 schools per district. To accommodate a request by Nepalese officials, the largest district, Morang district, was allocated a "double" sample of 24 schools. This increased the target sample size to 204 schools. In practice, the sample included only 203 schools, because the Solukhumbu district population of schools included only 3 non-priority schools. In what follows we describe sampling and weight calculations for the districts for which we sampled 12 schools; the same procedure was followed for Morang district, but all numbers in the following are doubled.

To reduce the probability of spillover from treatment to control schools, we chose the sample in a way that would reduce the probability than any given school in the sample was geographically proximate to any other school in the sample. This was done by randomly drawing small geographic areas called Village Development Committees (VDCs) and then drawing only one school per VDC. In the 16 districts from which the schools were drawn, there were 1,251 "eligible" schools (with at least grades 1 through 10 that had not received SSDP training) spread over 751 VDCs, so the average VDC had 1.67 eligible schools.

Within each district, we first sampled non-priority schools, by randomly sampling VDCs from among VDCs where any non-priority schools were located, with probability proportional to the number of such schools in the VDC, and then randomly sampling one non-priority school per selected VDC (if the VDC had more than one non-priority school). We then sampled priority schools by selecting VDCs from among those that had not been selected for inclusion in the non-priority stratum. Because some of the excluded VDCs included non-priority as well as priority schools, and because VDCs that include both non-priority and priority schools might be systematically different from VDCs including only priority schools, we increased the weight for the VDCs that included both non-priority and priority schools at this stage, with the aim of rendering the priority stratum sub-sample closer to representative of all priority schools within the district (see Schaffner, Glewwe and Sharma, 2018, for details). Aiming for an ultimate sample of four schools within the non-priority stratum and 8 schools within the priority stratum, we sampled six non-priority schools and 15 priority schools, so as to have two (seven) backup non-priority (priority) schools.[42] All

[42] In fact, during the actual data collection there were very few cases where the fifth or sixth non-priority school was included in the sample due to problems collecting data for the first four selected non-priority schools. Similarly, there were very few cases where the ninth or higher selected priority school was included in the sample because in almost all districts there were no problems collecting data from the first eight selected priority schools. Note that the selection of these "spare" schools, whether used or not, does not change the probability that any given school is selected into the sample; that is, selecting additional schools that may be used in the sample does not affect the probability that the first four or eight schools were selected into the sample.

sampling was done without replacement. More specifically, for selection within each stratum, all eligible VDCs were put into a list, with each VDC given as many rows in the list as there are eligible schools in the VDC. (A VDC's probability of selection depended on the number of lines it takes up in the list but not the order in which it appears in the list.) A random number was then selected for identifying the first line in this list to include in the sample, and then the sampling was completed by selecting lines at equal intervals down the list. The VDCs associated with the schools in the selected lines constituted the sample of VDCs. Unweighted random sampling was then used to select one eligible school per VDC (if the VDC had more than one eligible school). Further details are provided in Schaffner, Glewwe and Sharma (2018).

II. Procedure for Calculating Weights

Our aim was to create school-level population weights equal to the inverse of schools' probabilities of selection into the sample (from the population of schools with grades 1 to 10 in the 16 districts). Given the complicated process of sampling without replacement described above, we chose to calculate those probabilities using Monte Carlo methods. More specifically, we repeated the above process for sampling schools 10,000 times, each time randomly selecting a new set of starting points for selection of VDCs within the ordered lists of schools from eligible VDCs. For each VDC, the probability selection was calculated as the fraction of the 10,000 draws that the VDC was selected. For any given priority (non-priority) school, the probability of selection into the sample was the probability of its VDC being selected into the sample divided by the number of eligible priority (non-priority) schools in that VDC.

Appendix E: Two qualitative studies

Qualitative and phone interview report.zip

Appendix F: Supplementary tables

Supplementary table 1: Average endline test scores by student gender and ethnicity

Group (Sample sizes for math and science score average calculations)	Mean (std. dev.) endline math assessment percentage score	Mean (std. dev.) endline science assessment percentage score
Grade 9 – Males(n=2857,2857)	29.82 (9.34)	30.65 (9.23)
Brahmin and Chhetri(n=1102,1102)	30.57 (9.62)	32.10 (9.39)
Terai and Madheshi(n=322,322)	29.23 (11.06)	28.57 (10.79)
Dalit(n=344,344)	28.67 (9.01)	29.11 (8.31)
Newar(n=78,78)	29.07 (8.62)	29.85 (7.44)
Other Janajati(n=936,936)	29.36 (8.49)	30.06 (8.65)
Muslim(n=56,56)	25.73 (11.89)	26.37 (11.13)
Grade 9 – Females(n=3847,3847)	26.85 (9.31)	27.84 (8.53)
Brahmin and Chhetri(n=1391,1391)	27.42 (9.72)	28.27 (8.82)
Terai and Madheshi(n=397,397)	24.48 (10.39)	25.15 (9.86)
Dalit(n=480,480)	26.00 (9.05)	26.64 (7.65)
Newar(n=104,104)	28.92 (7.79)	29.83 (8.20)
Other Janajati(n=1423,1423)	26.66 (8.72)	28.01 (8.06)
Muslim (n= 26,26)	26.04 (11.83)	25.43 (12.03)
Grade 10 – Males (n = 2567,2568)	30.28 (9.73)	34.13 (9.87)
Brahmin and Chhetri(n =1025,1026)	31.17 (9.71)	35.24 (9.52)
Terai and Madheshi(n=264,264)	28.86 (12.10)	31.48 (11.86)
Dalit(n=341,341)	27.83 (9.35)	31.66 (9.22)
Newar(n=63,63)	34.14 (9.26)	36.09 (8.93)
Other Janajati(n=814,814)	30.27 (9.12)	34.12 (9.95)
Muslim(n=36,36)	28.63 (10.84)	33.22 (11.39)
Grade 10 – Females(n=3265,3265)	26.53 (9.34)	30.67 (9.00)
Brahmin and Chhetri(n=1339,1339)	26.76 (9.39)	31.26 (9.06)
Terai and Madheshi(n=304,304)	22.99 (10.88)	27.24 (9.62)
Dalit(n=364,364)	24.63 (8.49)	28.77 (8.06)
Newar(n=92,92)	29.10 (8.56)	32.79 (7.75)
Other Janajati(n=1118,1118)	27.28 (9.01)	31.03 (8.84)
Muslim(n=21,21)	24.73 (14.05)	30.23 (14.87)

Notes: Test score means and standard deviations are weighted.

Supplementary table 2: ITT estimates of impact of SSDP training on students' normalized test scores: full sample, all items: robustness checks

	Grade 9		Grade 10	
	Mathematics	**Science**	**Mathematics**	**Science**
Without weights				
Treat	-0.082*** (0.021)	-0.089***(0.022)	-0.075***(0.024)	-0.050** (0.025)
R^2	0.212	0.160	0.249	0.197
Sample size	6,800	6,797	5,832	5,829
Without controls for test-taking conditions				
Treat	-0.110 (0.066)	-0.111 (0.061)	-0.046 (0.073)	0.007 (0.074)
R^2	0.228	0.158	0.251	0.181
Sample size	6,800	6,797	5,832	5,829
Adding controls for variables that were not balanced at baseline				
Treat	-0.114(0.063)	-0.104*(0.058)	0.005 (0.073)	0.059 (0.071)
R^2	0.243	0.176	0.276	0.211
Sample size	6,800	6,797	5,832	5,829
Adding school, teacher and student controls				
Treat	-0.107(0.065)	-0.085 (0.063)	-0.063 (0.063)	-0.014 (0.084)
R^2	0.288	0.201	0.281	0.206
Sample size	5,658	5,383	4,752	4,602
Raw scores (percent of questions correctly answered) as dependent variable				
Treat	0.114* (0.067)	0.121** (0.057)	0.032 (0.070)	-0.033 (0.075)
R^2	0.218	0.150	0.250	0.175
Sample size	6,800	6,797	5,832	5,829

Notes: Estimates of β_T. The "default" estimates underlying the results in this table are those given in Table 5, which are 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). Standard errors, in parentheses, account for random assignment within strata and are clustered at the school level. Each panel of this table reports on regressions that depart from this default in just one dimension. The first panel uses Ordinary Least Squares rather than Weighted Least Squares estimation. The second omits the test-taking condition controls. The third adds school variables that were not balanced at baseline: log of total number of students, whether the school had electricity, whether the head teacher had a masters degree, and the percentage of the school's teachers who are female. The fourth adds the following variables: father had at least secondary education; mother had at least secondary education; an index of family assets; dummy variables for whether the teacher had SSRP training, had a permanent position, or less than five years of experience; the time it takes to walk from the school to the nearest all-weather motorable road (in indicator of remoteness). The fifth tests for impact on raw scores of students i.e. percent of questions correctly answered. Estimates that are statistically significant at the .10, .05 and .01 levels are indicated by *, ** and ***, respectively.

Supplementary table 3: Costs for SSDP trainings of 9th and 10th grade math and science teachers

Activity	Approach to estimation	Total in U.S. dollar equivalents for study roll-out[1]	Cost per ETC training session[2]	Cost per teacher[3]
Category 1: Costs incurred once for entire country				
Curriculum development and writing of guidelines	2 months of time at Under Secretary level per subject * 2 subjects (math and science)	1244.40	44.4	2.2
Category 2: Costs incurred once per ETC region				
Selection and training of ETC personnel, creation of trainer roster	This was not done afresh for this program, and these personnel and rosters are shared across many government teacher training programs.	We cost this at zero for our study intervention, but note that these costs might be required for replication in other contexts.	0	0
Training of trainers	In practice, there was no new training of trainers for this program, but the program was rolled out by ETCs in which trainers had received related training of trainers for related training programs in previous years.	We cost this as zero, but note that replication elsewhere would require training of trainers.	0	0
Category 3: Costs incurred once per training session				
Central oversight of financing arrangements	One month of time at Under Secretary level. Ministry of Education, Science and Technology approves the program and budget for teacher training and NCED gives permission to release funds (AKHATHARI) to related ETCs. This estimate is for the cost of administering funds for all 28 trainings in our study. Similar cost would be incurred again for additional waves of training sessions.	311.1	11.1	0.6
Selection of trainers by ETC	One week of time at Section Officer level * 14 ETCs * 2 trainings (one math and one science) per ETC	1940.9	69.3	3.5

Invitation of teachers, logistics, clerical assistance	Two weeks of time of a cleric* 14 ETCs * 2 trainings (one math and one science) per ETC	2821.6	100.8	5.0
Trainer time for preparation and delivery	Three weeks of time at ETC Subject Technical Officer levell* 14 ETCs * 2 trainings (one math and one science) per ETC	5822.7	208.0	10.4
Use of facilities	$136.63 * 14 ETCs * 2 trainings (one in math and one in science) per ETC; Estimate based on rental cost for 10 days in budget conference space in a district headquarter (NRs. 1500 per day)	3818.2	136.4	6.8
Participant per diems and lodging	$1768.50*14 ETCs * 2 trainings (one in math and one in science) per ETC; Based on actual cost from an example district.	49519.3	1768.5	88.4
Materials and handouts	$272.70 * 14 ETCs * 2 trainings (one in math and one in science) per ETC; Based on actual cost from an example district.	7636.4	272.7	13.6
Training follow-up by ETC personnel	In practice, this did not occur.	We are costing this at zero, but note that in principle this should have taken place, and might be valuable if replicated elsewhere.	0	0
Totals		73114.6	2611.2	130.6

Notes: [1] The exchange rate used is US$ 1 =NRs. 110. [2] Where relevant, totals are divided across 28 sessions (2 sessions each, one in math and one in science, in each of the 14 ETCs relevant to our 16 study districts).[3] We assume an average of 20 teachers per ETC training session.

References

Abeberese, AB, Kumler, TJ and Linden, LL, 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), p611-633.

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, F, Anauati, MV, Furman, M, Luzuriaga, M, Podestá, ME and Taylor, I, 2019. 'Training to Teach Science: Experimental Evidence from Argentina', *World Bank Economic Review*, 34(2), p393-417.

Araujo, MC, Carneiro, P, Cruz-Aguavo, Y and Schady, N, 2016. 'Teacher Quality and Learning Outcomes in Kindergarten', *Quarterly Journal of Economics*, 131(3), p1415-1453.

Banerjee, A, Cole, S, Duflo, E and Linden, L, 2007. 'Remedying Education: Evidence from Two Randomized Experiments in India*', Quarterly Journal of Economics*, 122(3), p1235-1264.

Benjamini, Y and Hochberg, Y, 1995. 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), p289-300.

Benjamini, Y and Yekutieli, D, 2001. 'The Control of the False Discovery Rate in Multiple Testing Under Dependency', *Annals of Statistics*, 29(4), p1165-1188.

Berlinski, S and Busso, M, 2013. 'Pedagogical Change in Mathematics Teaching: Evidence from a Randomized Control Trial', Washington, DC: Inter-American Development Bank.

Bruns, B, Costa, L and Cunha, N, 2018. 'Through The Looking Glass: Can Classroom Observation and Coaching Improve Teacher Performance in Brazil?', *Economics of Education Review*, 64, p214–250.

Bruns, B and Luque, J, 2015. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: The World Bank.

Chetty, R, Friedman, J and Rockoff, J, 2014. 'Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates', *American Economic Review*, 104(9), p2593-2632.

Cilliers, J, Fleisch, B, Prinsloo, C and Taylor, S, 2018. 'How to improve teaching practice? Experimental Comparison of Centralized Training and In-classroom Coaching'*.* RISE Working Paper 18/024. Research On Improving Systems of Education (RISE).

Conn, KM, 2017. 'Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations'*, Review of Educational Research*, 87(5), p863-98.

Damon, A, Glewwe, P, Wisniewski, S and Sun, B, 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.

Duflo, E, Dupas, P , and Kremer, M, 2011. 'Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya', *American Economic Review*, 101(5) p1739-74.

Evans, DK and Popova, A, 2016. 'What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews' *World Bank Research Observer*, 31(2), p242–270.

Finan, F, Olken, B and Pande, R, 2015. 'The Personnel Economics of the State'*.* NBER Working Paper Number 21825. Cambride, MA:National Bureau of Economic Research.

Fuje, H, and Tandon, P, 2018. 'When Do In-service Teacher Training and Books Improve Student Achievement? Experimental Evidence from Mongolia'*, Review of Development Economics*, 22(3),1p360-1383.

Jukes, MC, Turner, EL, Dubeck, MM, Halliday, KE, Inyega, HN, Wolf, S, Zuilkowski, SS and Brooker, SJ, 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), p449-481.

Kafle, B, Acharya, SP and Acharya, D, 2019. *National Assessment of Student Achievement 2018: Main Report.* Bhaktapur: Education Review Office (ERO), Ministry of Education, Science and Technology. Kathmadu, Nepal.

Loyalka, P, Popova, A, Li, G and Shi, Z, 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.

Lu, M, Loyalka, P, Shi, Y, Chang, F, Liu, C and Rozelle, S, 2019. 'The Impact of Teacher Professional Development Programs on Student Achievement in Rural China: Evidence from Shaanxi Province' *Journal of Development Effectiveness*, 11(2), p105-131.

Lucas, AM, McEwan, PJ, Ngware, M and Oketch, M, 2014. 'Improving Early Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda', *Journal of Policy Analysis and Management*, 33(4), p950–976.

Macdonald, K, and Vu, BT, 2018. 'A Randomized Evaluation of a Low-Cost and Highly Scripted Teaching Method to Improve Basic Early Grade Reading Skills in Papua New Guinea', Policy Research Working Paper 8682. Washington, DC: The World Bank.

Macdonald, K, Brinkman, S, Jarvie, W, Machuca-Sierra, M, Mcdonall, K, Messaoud-Galusi, S, Tapueluelu, S and Vu, B, 2018. 'Intervening at Home and Then at School: A Randomized Evaluation of Two Approaches to Improve Early Educational Outcomes in Tonga', Policy Research Working Paper 8427. Washington, DC: The World Bank.

McEwan, PJ, 2015. 'Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments', *Review of Educational Research*, 85(3), p353 394.

Ministry of Education, 2016. *School Sector Development Plan 2016–2023.* Kathmandu: Ministry of Education, Government of Nepal.

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

National Centre for Educational Development (NCED), 2017a. *Secondary Level TPD Training Curriculum: Science.* Bhaktapur: Ministry of Education, Government of Nepal.

National Centre for Educational Development (NCED), 2017b. *Secondary Level TPD Training Curriculum: Mathematics.* Bhaktapur: Ministry of Education, Government of Nepal.

Piper, B, Zuilkowski, S, Dubeck, M, Jepkemei, E and King, S, 2018. 'Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides', *World Development* 106, p.324-336.

Popova, A, Evans, D and Arancibia, V, 2016. 'Training Teachers on the Job: What Works and How to Measure It', Policy Research Working Paper 7834. Washington, DC: The World Bank.

Popova, A, Evans, D, Breeding, M and Arancibia, V, 2019. 'Teacher Professional Development around the World: The Gap between Evidence and Practice', Working Paper 517. Washington, DC: Center for Global Development.

Powell, D, 2017. *Quantile Treatment Effects in the Presence of Covariates*. Santa Monica, CA: RAND Corporation.

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, S, Hanushek, E and Kain, J, 2005. 'Teachers, Schools, and Academic Achievement', *Econometrica* 73(2), p417-458.

Schaffner, J, Glewwe, P and Sharma, U, 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, J, Glewwe, P and Sharma, U, 2019a. '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, J, Sharma, U and Glewwe, P, 2019b. 'Federalism in Nepal: Early Implications for Service Delivery in Education', Unpublished.

Shrestha, D, 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.

Snilstveit, 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).

Westbrook, J, Durrani, N, Brown, R, Orr, D, Pryor, J, Boddy, J and Salvi, F, 2013. *Pedagogy, Curriculum, Teaching Practices and Teacher Education in Developing Countries.* Final Report. Education Rigorous Literature Review. London: Department for International Development.