Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru

Stephanie Majerowicz and Ricardo Montero[*]

November 15, 2018

Job Market Paper
Please find the latest version here.

Abstract

Despite the massive worldwide increase in school enrollment over the past 50 years, many students are not learning. Teacher quality—a key determinant of student achievement—remains low in many countries, and while governments invest heavily on teacher training programs, the evidence on their impact is inconclusive. Development economists, skeptical of teacher training, have instead largely focused on incentive programs. We present novel evidence on the impact of a national teacher coaching program in Peru using an at-scale randomized controlled trial. The program provided teachers in rural primary schools with individualized, continuous coaching on pedagogical practices. Coaching substantially improved learning: students in treatment schools experienced a 0.25-0.38 standard deviation increase in standardized test scores relative to the control group. The gains are observed throughout the test score distribution, with low-performing students benefitting as much as higher performing ones. Using a combination of experimental and non-experimental techniques to account for teacher rotation, we show that the program effects persist for at least one year after the training ends. Interestingly, the impact observed is entirely due to and retained by the trained teacher—schools that lose trained teachers lose the entire initial gains and, when treated teachers move, students benefit from the arrival of the trained teacher as much as students in the original school did. This suggests that the program is building up the human capital of teachers, rather than simply monitoring teacher presence or effort, and that this human capital is portable and persistent. Our results have important policy implications. We find that teacher training programs can indeed be impactful and cost-effective but, given the high level of teacher movement, individual schools may underinvest in teacher training thereby underscoring the need for public subsidies of such training.

[*] Majerowicz: Harvard University. Email: smajerowicz@fas.harvard.edu; Montero: University of Minnesota. Email: monte290@umn.edu. Majerowicz is deeply grateful to her advisors Asim Khwaja, Rema Hanna and Michael Kremer for their guidance throughout this project. This project would not have been possible without the support of María Antonieta Alva, José Carlos Chávez and Minister Jaime Saavedra in the Ministry of Education of Peru, as well as the assistance of Ministry staff, particularly Andrei Bartra, María Pia Basurto, Fabiola Caceres, Antonio Campos, Walter Carrillo, Giuliana Espinoza, Harold Galvez, Liliana Miranda, and María Alejandra Whang. We are also particularly grateful to Felipe Barrera, Juan Francisco Castro, Juan Sebastián Galán, Siddarth George, Paul Glewwe, Lorena Caro, Soeren Henn, Joaquin Klott, Mateo Montenegro, Juan Felipe Olano, Nicolás Riveros, los Mellos Romero, Román Andrés Zárate and participants at the Harvard Economic Development Seminar for helpful comments on the paper.

1 Introduction

There is robust evidence that teacher quality is a fundamental determinant of student outcomes (Chetty et al., 2014), but much less is known about how to improve teacher skills. Most developing countries spend significant amounts of their education budget on teacher training programs aimed at improving the skills of the current stock of teachers (Bruns and Luque, 2014). However, the effectiveness of in-service training in improving teacher performance continues to be an area of great debate, and the large literature on teacher training is inconclusive (World Bank, 2018). The economics literature has traditionally been skeptical of teacher training, favoring instead incentive systems and monitoring mechanisms like pay-for-performance (Muralidharan and Sundararaman, 2011), monitoring teachers with cameras (Duflo et al., 2012), and short contracts that elicit higher effort from teachers (Duflo et al., 2015; Muralidharan and Sundararaman, 2013).[1] The education literature, on the other hand, generally views teacher training more favorably, but rigorous evidence of specific programs is mixed with as many positive as null results on student learning outcomes (Hill et al., 2018; Villegas-Reimers, 2003).[2]

Adding to the confusion, and probably contributing to the mixed results in the literature, is the fact that teacher training can mean any number of different things: from single theoretical workshops, to peer-to-peer engagement, to multi-year coaching programs. Furthermore, teacher training can focus on upgrading subject-specific knowledge or on improving pedagogical practices. These skills may have strong complementarities, so the effect of improving one margin will depend heavily on level of the other one, which means teacher training programs may have heterogeneous effects by initial teacher characteristics as well as by complementary school inputs. Third, there is a wide variation in the scale and quality of the programs, as well as in the scale and quality of the evaluations.[3] For all of these reasons, it is perhaps not surprising that the training literature is inconclusive.

[1] See Glewwe and Muralidharan (2015), Ganimian and Murnane (2016) and Kremer et al. (2013) for recent reviews of this education literature in economics. This traditional focus on incentives has recently started to change, led by a recent review by Evans and Popova (2016) that points to some successful studies that involve teacher training alongside materials and technology, but that had previously been categorized as other types of interventions.

[2] While the education literature has consistently argued for the importance of professional development for teachers (Bal and Cohen, 1999; Schleicher, 2016), results of individual evaluations have been mixed. Early studies showed positive effects, while more recent evaluations mostly find no impact. Of 90 studies funded by the Institute for Education Sciences since 2002, 88% produced weak or null results (Coalition for Evidence-Based Policy, 2013).

[3] Yoon et al. (2007) find that only 9 out of 1,300 studies they review meet What Works Clearinghouse evidence standards.

In this paper, we focus on one type of in-service training program, known as coaching, in which experienced tutors are sent to observe teachers and provide feedback on pedagogical practices on a regular basis. Institutions like the World Bank (2018) are promoting teacher training programs like these that provide ongoing support to teachers and focus on practical classroom skills, rather than one-off, short, theoretical workshops that have not proven to be very effective. Coaching programs have three key characteristics. They are: (1) individualized – they provide one-one sessions with teachers tailored to their individual needs, (2) sustained – they take place over a long period of time like a year or more, and (3) intensive – sessions occur regularly every few weeks or every month.

In 2010, Peru began implementing a national coaching program that sends experienced tutors or "coaches" to visit teachers once a month. During each visit, tutors spend an entire day with each teacher that includes 5 hours of classroom observation and 3 hours of feedback focused on pedagogical practices. The program currently reaches over 1.5 million students and 73,000 teachers in 13,000 schools, and costs the government 130 million dollars per year. Besides Peru, these coaching programs are also being implemented in a number of Latin American countries including in Colombia, Brazil and Chile, also at significant cost, yet their impact on the learning outcomes of students has not been rigorously evaluated.

This paper takes advantage of a national experiment in Peru that randomly assigned the teacher coaching program to over 6,000 rural primary schools in order to estimate its impact on student learning outcomes. The experiment also randomized the removal of the program from a subset of the schools that had been receiving it, allowing us to explore its persistence and the mechanisms through which the program operates.

Coaching substantially improved learning: the coaching program had large, positive effects of between 0.25 and 0.38 SD on student learning outcomes measured in second grade standardized test scores. The effects are constant throughout the ability distribution, which suggests that all students are benefitting from the program. We find no evidence of impact on students' grades as assessed by their teachers or on students' grade repetition or drop outs. This suggests that teachers adjust their grading criteria in the classroom to the average of their students rather than the expected learning outcomes of the grade, consistent with the fact that the entire student distribution is shifting to the right.

The program is designed to build up the human capital of teachers. However, it is possible that the tutors who show up to distant schools that otherwise receive little to no attention from the local or national governments could be serving the function of monitors, incentivizing teachers to show up and to exert more effort while the tutor is present in the classroom. We can distinguish between these two channels by looking at the persistence of the treatment effects after the program ends. We expect that if the treatment is providing teachers with new pedagogical skills, the treatment effect will persist as teachers are left with new classroom tools. However, if the program is functioning through monitoring, its effects should disappear once the "monitors" leave.

In order to distinguish between these two channels of skill development and monitoring, we take advantage of the second experiment that randomly removes the program from a subset of schools that had been receiving it.

However, given the high level of teacher rotation in Peru — each year 30% of Peruvian teachers move to new schools — we must take into account teacher rotation in order to test the persistence of the treatment once the program ends distinguishing between those schools that keep the trained teachers and those that do not. In order to be able to take into account teacher rotation, we first test whether the program is either having a direct effect on teacher rotation, or is associated with characteristics that make teachers more likely to move. We first test the effects of the random program assignment in 2016 on teacher rotation in 2017, and find that the program does not impact teacher rotation. We then use a simple machine learning lasso algorithm for model selection to identify those characteristics that are most predictive of moving. We find that while teachers who move are different from those who stay in important ways, the treatment does not interact with these variables in ways that could bias our estimator. This allows us to test what happens when schools lose the program, conditional on keeping their teachers.

Once we take into account high teacher rotation, the program effects persists in schools that lose the program as long as the treated teachers remain. These results suggest that the program is in fact building up teachers' human capital rather than operating through a monitoring effect. At the same time, schools that lose the program and lose their treated teachers experience a large drop in test scores relative to schools that keep the program equal to the entire magnitude of the program gains. This suggests that the coaching works exclusively through the teacher and is not having spillover effects on the rest of the school (principal, parents, or students) that could persist once the treated teachers leave.

We follow treated teachers who move to non-treated schools in order to test the persistence of the program outside the original treated school. While we would like to use our experimental sample to follow randomly treated teachers, outcome data for 2017 are unavailable. However, using a difference-in-difference estimator on an earlier sample of treated schools, we are able to follow treated teachers to non-treated schools and find that the entire effect of the program persists one year later. In other words, students in the schools that receive the treated teachers benefit as much from the program as students in the original school did. This confirms the idea that the program is working directly through building up teachers' human capital, and that teachers retain the full effects of the program even when they move to new schools. The program is therefore not only effective at improving test scores, but its effects are also highly persistent.

However, the high mobility of teachers suggests that individual schools are likely to underinvest in teacher training given the likelihood that teachers leave taking their human capital investment with them. This is similar to findings in the labor and public finance literature that firms will underinvest in general worker training if workers are mobile (Becker, 1962). This implies that the government has a key role in providing or subsidizing teacher training taking into account positive externalities to schools that receive trained teachers. This also has implications for policies that decentralize education spending to local governments or schools, which could lead to underinvestment in programs that target the human capital formation of teachers. In Peru, for example, most of the teacher mobility occurs within the regions (10% of teachers move across school districts, but only 2% across regions) so it may make sense to decentralize education spending to regional governments, but perhaps not to districts or schools. Similarly, education policies that result in substantial privatization of school systems may lead to suboptimal teacher training, unless schools find ways to retain teachers or governments subsidize trainings.

We explore heterogeneous impacts by various teacher characteristics. The program works best for younger teachers, which suggests that older teachers are perhaps less open to implementing some of the new techniques or find older habits harder to break. While it could be that the coaching program is accelerating the learning curve that occurs with experience, we do not find differential impacts by measured experience. There are also no differential effects by the job security of teachers, which we would have expected if the program had been working as a monitoring tool. Finally, teachers with higher initial cognitive skills and with higher content knowledge of their teaching area benefit the most from this program. This suggests strong potential complementarities between the pedagogical training provided by the treatment and other policies intended to improve either teacher selection into the profession (for example through competitive pay) or strengthen content knowledge. This heterogeneity also sheds light on the mixed findings of the general training literature, which may be explained in part due to variation in complementary teacher skills.

We calculate the cost-benefit analysis of this program under various assumptions of the decay of the treatment effect and of the duration of the training (whether teachers are trained for one or two years). We make assumptions about two sources of program decay: teacher exit from the school system which we calibrate using our administrative teacher records to be between 5-7% per year, and natural decay of program effects as training starts to wear off or becomes obsolete. While the program is expensive, once the persistence of the program is taken into account, we find that the program becomes relatively cost-effective with benefits between 0.72 and 1.12 SD per 100 dollar investment for 5-10% annual decay. Training programs can therefore be a relatively cost-effective mechanism to improve teacher quality, particularly compared with incentive schemes that raise test scores, albeit at a significant and recurring cost to the government.

This paper makes contributions to the literature on teacher training, specifically on coaching, as well as to the evidence base of large-scale experimental government evaluations.

Most of the evidence of the impact of coaching programs comes from small pilot programs in developed countries (Kraft et al., 2018; Allen et al., 2011; Biancarosa et al., 2010; Matsumura et al., 2013, 2012; Campbell and Malkus, 2011). Three small RCTs in developing countries that provide coaching along with pedagogical materials find positive impact on teacher performance, and two find impact on student learning while the third does not (Albornoz et al., 2018; Cilliers and Taylor, 2017; Yoshikawa et al., 2015). However, it is impossible to disentangle the effect of coaching from the effect of the entire bundle of materials provided to teachers. The only evidence of a coaching program rolled out on a national scale comes from a non-experimental evaluation Colombia's "Programa Todos a Aprender," which provides schools with textbooks, teacher coaching and principal training. Using a regression discontinuity design, Barrera-Osorio et al. (2018) find no significant impact on learning outcomes, and attribute it to problems with the design and implementation of the program (there were few tutor visits and those visits lacked structure), although it could reflect a local average treatment effect of zero for schools at the cutoff.

We also contribute to the experimental evidence done at scale. There are general concerns about extrapolating from small RCTs to implementation at scale, and reasons why scale-up could be particularly challenging in coaching programs. Muralidharan and Niehaus (2017) highlight three: (1) issues that arise from scaling up interventions either due to reduced ability to monitor or due to changing from implementation by an NGO to the government, (2) study samples may not be representative of the population of interest, and (3) the experiment may not capture spillover or general equilibrium effects. The complexity of coaching programs, which require ensuring that high quality tutors regularly show up in distant schools, mean that they can be particularly challenging to implement successfully at national scale. The implementation issues raised by Barrera-Osorio et al. (2018) in Colombia provide some indication that ensuring faithful implementation can be challenging even for middle income countries. Similarly, we may be concerned that small scale training programs may have teachers that are not representative of the wider pool of public school teachers, affecting the external validity of these estimates (having above-average teachers in pilot programs may actually underestimate treatment effects).

This paper presents the first evidence of a randomized coaching program done at national scale implemented by any government. This allows us to address the three issues raised by Muralidharan and Niehaus (2017): first, we study a national program being implemented at scale by the government; second, schools (and teachers) in our sample are representative of the universe of schools for which the program is intended, and third, we are able to estimate spillovers by following treated teachers as they move into non-treated schools. We also contribute to the broader literature on teacher training by exploring channels of persistence and heterogeneity that shed light on the mixed findings in the literature.

The rest of the paper is organized as follows: Section 2 discusses the Peruvian school context and provides a description of the coaching program. Section 3 discusses the experimental design, the available data and the empirical strategy. Section 4 presents the main results. We explore channels and persistence in Section 5, and heterogeneity by various teacher characteristics in Section 6. Section 7 presents cost-benefit calculations, and Section 8 concludes.

2 Background

Despite important recent improvements, Peru has low educational performance relative to other middle-income countries as measured by international exams. In the 2015 PISA exams, Peru ranked 63 in Science, 61 in Mathematics, and 61 in Reading out of 69 participating countries, behind most of its peers in Latin America (OECD, 2016).

Studies of Peruvian teachers show large gaps on both content knowledge and pedagogical practices. A survey of sixth grade teachers found that 84 percent scored below level 2 in an exam testing sixth grade math skills, and almost 50 percent scored below level 2 in language skills (Bruns and Luque, 2014). Less than 3 percent reached level three in math and less than 25 percent in language, a level that implies mastery of the content they are supposed to teach.

Peruvian teachers also perform poorly in measures of pedagogical practices. For example, Peruvian teachers spend only 60 percent of their class time on academic activities, and around 13 percent off-task, relative to best practices of over 85 percent and zero respectively (Bruns and Luque, 2014).

2.1 Program Description

In order to address these gaps in teacher skills, in 2010 Peru began to implement a coaching program that sought to improve the pedagogical practices of teachers.

The program, "Acompañamiento Pedagógico Multigrado" is an in-service training program in which experienced teachers (called tutors) are sent to provide support and give feedback to teachers in rural, primary schools in Peru. The program has two main components: monthly, individual classroom visits, and monthly group training workshops. Each tutor completes 9 classroom visits to each teacher (1 diagnostic session, 7 for general observation, training and feedback, and a closing session), and 8 micro-workshops per year.

During each visit, tutors spend an entire day with each teacher, which includes 5 hours of classroom observation and 3 hours of feedback during which the tutor reviews her observations with the teacher, discusses mistakes and areas of improvement, and practices pedagogical tools with the teacher. The first session is a diagnostic session in which the tutor observes and grades the teacher on a rubric according to her performance on pedagogical practices, and draws up a coaching plan for the year in order to improve the teacher's pedagogical skills. In addition to the classroom observations, each tutor performs 8 workshops with all the teachers in his/her charge to discuss pedagogical practices and encourage the exchange of ideas.

Instead of content knowledge of the material, the program focuses on strengthening pedagogical practices and on developing the ability of teachers to reflect on their own strengths and weaknesses and adjust their behavior accordingly on a constant basis:

> "The pedagogical accompaniment promotes the development and strengthening of the skills related to understanding the student in her context, curricular planning, guiding the learning, school environment, and evaluating student learning. In addition, it promotes the development of critical thinking skills like auto-reflection and analysis, through exercises that seek reflection and critical analysis of the teacher's performance." (Ministerio de Educación de Perú, 2016).

This program works in a cascade system, with each tutor in turn trained, supported and monitored by a *Pedagogical Specialist.* Each *specialist* is responsible for visiting/monitoring each tutor at least twice a year during his or her classroom visits. In addition, the *specialist* provides two workshops directly to teachers per year.

While the program is designed by the Ministry, it is implemented by each local school board (UGEL). Tutors are hired by the local school board from the current teacher pool and trained by the specialists. Tutors are supposed to be exceptional teachers: to be eligible to be tutors, teachers must hold a pedagogical university degree, have at least 5 years of teaching experience, as well as 1-2 years of experience in training or providing support for teachers. In 2016, tutors earned 3,600 soles per month (roughly 1,200 US dollars), much more than regular teachers who would earn a starting salary of 1,500 soles, and could expect to reach a salary of 3,000 soles at most (the average teacher salary in 2016 was of 1,850 soles or 600 USD).

There are three versions of this coaching program that target different subsets of schools in the Peruvian education system. We focus on the program that targets poorly performing rural *multigrade* primary schools.[4] Multigrade schools are schools in which there are fewer

[4] Altogether, the coaching programs reach over 1.5 million students in 13,000 schools and over 73,000 teachers. Two other programs implement the coaching intervention in other target populations:(1) *Asistentes de Soporte Pedagógico Intercultural*, targets *bilingual* schools (schools where a significant proportion of students do not speak Spanish at home). The tutors are therefore required to be fluent in one of the local languages spoken, and schools are grouped in "networks" that are assigned to various tutors. (2)

teachers than grade levels, and they present particularly difficult teaching challenges since teachers have to accommodate different grade levels in the same classroom. Multigrade schools are small, rural schools with 30 students and 2 teachers on average, and they represent the largest number of schools with coaching programs.

The rural multigrade program is particularly expensive because the target population schools tend to be fairly distant from one another, which means that the program requires a large number of tutors and significant travel expenses. The coaching program in multigrade schools alone cost the government approximately 40 million USD in 2016, and benefitted 174,000 students (see Table A.1 in the Appendix). This translates into an annual cost of 228 USD per student, which is more than 20% of the total expenditure per student in primary school (for 2015 the average spending for primary school students was 2,800 soles or approximately 940 USD).

2.2 Teacher Rotation

Teacher rotation in Peru is fairly high, with approximately one-third of all teachers switching schools at the end of each school year. This is in part explained by the fact that 40% of Peruvian teachers are contract teachers, whose contracts expire each year and a significant portion of those teachers move to a new school. While it is less common for a teacher to move during the school year, it is also possible, and in this case tutors would switch to providing the coaching to the new or remaining school teachers, rather than follow the teacher to his or her new school. We will need to take into account this high teacher mobility because we risk underestimating the treatment effects and particularly the persistence of the treatment when teachers move.

Soporte Pedagógico targets urban primary schools that have at least one teacher per grade. Besides pedagogical coaching, this version of the program includes the provision of materials, remedial classes, and support in school management practices. This program has also been rolled out in secondary schools that have started to implement an extended school day (*Jornada Escolar Completa*).

3 Data and Empirical Strategy

3.1 Experimental Design

The Ministry of Education randomized a re-assignment of the coaching program to new schools starting in 2016 in order to facilitate its rigorous evaluation. The expansion of the program had to meet certain regional quotas, which took into account the ability of each region to hire enough qualified tutors to cover the new schools. As a result, the randomization was stratified by region. Table A.2 shows the number of control and treatment schools by region.

The randomization was done by first restricting the universe of primary schools in Peru to those that met the eligibility criteria: being monolingual (Spanish speaking), multi-grade, and having low standardized test scores. Schools that met the criteria were selected for the sample and randomly assigned to a treatment and control group by region, provided there were more available schools than the quotas in program expansion for those regions. In total there are 6,207 schools in the experimental sample.

Schools that had received one or two years of coaching were automatically selected to continue in the program (and are not part of the evaluation sample), but those schools that already had three years of treatment and should have "graduated" from the program were randomly selected into the treatment and control groups. This resulted in two separate experiments:

1. *The random assignment* of the program for 4,526 schools not currently receiving the program. Of these, 1,922 were assigned to the control group and 2,604 to the treatment group.
2. *The random removal* of the program for the 1,678 schools that had been receiving the program between 2013-2015. Of these, 1,186 were assigned to keep the program, and 492 to have it removed.

These schools were selected for the program in 2015, and began receiving the program starting in 2016. The school year begins in March, and the standardized test scores are taken in November so that students and teachers would have participated in the program for one full academic year by the time they take the standardized test scores.

In addition to these two experiments in which the treatment effects are identified due to random assignment, to test some of the channels and the persistence of the treatment as teachers move, we will also be relying on an earlier sample of treated schools using a difference-in-difference identification strategy. This second identification strategy, discussed in more detail in section 5, takes advantage of the availability panel data for earlier cohorts of treated schools and follow treated teachers as they move to non-treated schools.

3.2 Empirical Strategy

Given the fact that the intervention was randomized, the average treatment effect is given by a simple difference of means. Since the randomization was stratified by region, we include region fixed effects in all specifications. As school boards (UGEL) were responsible of the actual implementation of the intervention, we also include school district fixed effects in some specifications to control for variations in the implementation of the program within each region. Our main specification is the following:

$$ECE_s = \alpha + \beta \text{ Treat}_s + \lambda_r + \mathbf{X'}\Gamma + \varepsilon \tag{1}$$

Where ECE is the score of school s on the standardized exam, Treat is a treatment dummy, and λ_r is a region or school district fixed effect. We include a vector $\mathbf{X}$ of covariates to adjust for imbalance in baseline and for efficiency. For specifications where the unit of observation is the student and not the school, we cluster standard errors at the school level, which is the level of the treatment.

A second specification exploits the availability of panel data for schools to include school and year fixed effects for a period from 2007-2016:

$$ECE_{ts} = \beta \text{ Treat}_{ts} + \alpha_s + \delta_{tr} + \varepsilon_s \tag{2}$$

While school fixed effects are not necessary for identification given that the treatment was randomly assigned, they can help with precision. To account for variation in time trends by region, we include region-by-year fixed effects (δ_{tr}). Standard errors in this specification are also clustered at the school level.

3.3 Data

We use the following administrative datasets:

Evaluación Censal de Estudiantes (ECE): The primary measure of student learning outcomes is the standardized exam taken at the end of 2nd grade of primary school that tests mathematics and reading comprehension. The exam has been implemented each year since 2007,[5] and tests all schools with at least 5 students in second grade. Table A.3 in the Appendix shows descriptive statistics of the exam. The ECE scores are reported both as levels of subject mastery[6] and as a Rasch score with a mean of 500 and standard deviation of 100. Table A.3 shows that overall test scores are low across all schools, as a large proportion of students are ranked in the lowest category possible. For example, in 2015 only 22% of students met the expectations for learning in their grade in Mathematics. Test score data are available at the individual student level.

School Grades: Data on student grades and grade repetition was obtained from SIAGIE, a system that records enrollment, grades and the student results at the end of each year for all students since 2013. Students are evaluated by their teachers on a scale from 1 to 4, where 1 represents "beginning" comprehension of the material, 2 "in progress," 3 "satisfactory," and 4 "outstanding." Students pass a course with a grade of 3 or higher, and fail the year if they have an average score less than 3. Table A.3 in the Appendix shows the average grades for students in the experimental sample. It shows sharp differences between the assessment of students' performance by the teachers and the ECE.[7] While according to the ECE between 14% and 20% of students score a satisfactory level, teachers evaluate over 90% of students as "satisfactory" or "outstanding." Only between 6% and 9% of all students fail the year and have to repeat it.

School characteristics come from administrative datasets as well as an annual school census survey, *Censo Escolar.* Available characteristics include information on the number

[5] The standardized exam was continuously implemented from 2007 until 2016, it was discontinued in 2017 for one year due to a Ministerial decision. Previously that year, teachers went on prolonged nationwide strike and students had missed several weeks of class and were not up to date with the subjects that the ECE covered.

[6] There are three categories' 'beginning," "in progress", and "satisfactory."

[7] As both ECE and SIAGIE include a common student ID since 2014, it is possible to estimate how correlated both measurements are within each school. We recode grades in SIAGIE to match the ECE by combining the "satisfactory" and "outstanding" categories. In both 2014 and 2015, the average correlation between the ECE and classroom scores is very low, ranging from 12% to 14% for reading and between 7% and 11% for math.

of students, teachers, materials, school geo-coded location, and school infrastructure.

Teacher characteristics come from the administrative dataset (NEXUS) of all teachers and administrators in the public school system. It has information on teacher characteristics like age, education, and type of teacher contract. In addition, we have a measure of cognitive ability from the exam teachers take to get into civil service career, which tests cognitive skills in mathematics and language, and subject-specific knowledge of their teaching area and grade. These career entrance exams were given in 2009, 2010, 2011 and 2015, although the exam scores are not comparable across years.

Student Characteristics: We have additional information on individual student characteristics including student gender, and socioeconomic characteristics of the family.[8]

Region/UGEL Characteristics: Peru is divided into 25 regions (departments), which are further subdivided into 248 school districts managed by local school boards (UGEL, Unidad de Gestión de la Educación Local). We have characteristics at both region and school district levels, including district poverty, measures of institutional strength, indigenous population, etc.

Table 1 shows descriptive statistics some of these variables for schools in the experimental sample that took the standardized test in 2016.

3.4 Baseline Covariate Balance

To ensure that the randomization yielded a balanced treatment and control group we check the baseline balance on a number of school characteristics. Figure 1 and Table 1 show the coefficient point estimates for a number of baseline characteristics, which have been standardized for comparability.[9]

Figure 1 shows that most covariates are balanced, with the exception of number of students and, consequently, number of teachers since teachers are assigned as a function of the number of students. While this is roughly what we would expect by chance to be

[8] All data where merged and anonymized directly by the Ministry to protect student information.

[9] Given the randomization was stratified by region, the regressions in the Figure also include regional dummies, since we expect characteristics to be balanced within but perhaps not across regions. Baseline characteristics are equally balanced if we include school district fixed effects. Columns 3-5 of Table 1 shows the control and treatment means for the various covariates, which also gives us a sense of the characteristics of the sample.

significant, and the student-teacher ratio, which might affect treatment outcomes directly, is balanced, we control for these two variables in our regressions to ensure that we are not attributing any differences due to this size imbalance to our treatment. Figure A.1 in the Appendix shows the baseline covariates separately for the subsamples in the two experiments, which shows both experiments are balanced at baseline.

4 Results

4.1 Standardized Test Scores

The coaching program substantially improves learning. Table 2 shows the average treatment effect of the program on standardized test scores.[10] Columns 1 and 4 show the difference in means between the treatment and control groups, with fixed effects by region to account for the fact that the randomization was stratified by region.[11] Because the program was designed by the Ministry of Education but implemented by each local school board (UGEL),[12] Columns 2 and 5 include school district fixed effects that control for any differences in the actual implementation of the program within each region. Finally, Columns 3 and 7 take advantage of the availability of panel data for schools from 2007 to 2016, and include school level fixed effects along with year-by-region dummies to capture state-specific time trends. Because we have panel data, the number of observations is much larger in this specification, although we cluster standard errors by school, which is the unit of treatment. While the school fixed effects are not necessary for identification, they help with precision.

We find that the coaching program has a strong, positive impact on student learning. Average student test scores increase by 0.19 standard deviations in Math and 0.12 standard deviations in Reading Comprehension for the treated schools relative to the control group.

[10] The standardized test scores are only available for schools with more than 5 students in second grade, which reduces the sample of schools. Table A.2 shows, by region, the number of schools that have test scores in our treatment and control groups. The baseline balance in Figure 1, however, is done over this smaller subsample of schools which is the relevant sample for future analysis, with the exception of some outcome variables which we have for the entire sample.

[11] All of the specifications shown in Table 2 also control for school size which is slightly unbalanced at baseline, although the results are robust to excluding it.

[12] Peru has a total of 225 districts managed by school boards, which are the entities responsible for implementing education policies in the territory. Each UGEL is overseen by its Regional Education Board (Dirección de Educación Regional, DRE)

These results are identical for both of our preferred specifications (school district fixed effects and panel data), and suggest that coaching programs that provide regular, individualized support to teachers can be an effective mechanism to increase student learning. The program was highly effective despite the many implementation challenges that a complex program like this faces when implemented by local governments with institutional capacity constraints.

4.1.1 Adjusting for Teacher Rotation

Teacher rotation in Peru is fairly high, with approximately one-third of all teachers switching schools at the end of each school year. While our main results show the effect for a school of being randomly assigned the program in 2016, teacher rotation in previous years means that we have variation in the proportion of effectively treated teachers in each year. More specifically, while none of the schools were being treated in 2015, some teachers trained in schools treated in prior years moved to and were working in sample schools in 2016 affecting the proportion of treated teachers in both control and treatment schools.

While the movement of teachers is not random, we can use the random assignment in 2016 to instrument for the proportion of effectively treated teachers in each school. Since we have data on teachers' school location going back to 2015, we construct a variable that captures the proportion of treated teachers in two ways. First, we code a teacher as treated if she was treated in either 2015 or 2016, and a school as treated if the school received any treated teachers or was treated itself. This takes value of 1 for all schools in our treatment group by construction since all of them received treatment in 2016, but adjusts our earlier estimate for the fact that some control schools had received teachers who had been exposed to the program in prior years.

Second, we code the proportion of teachers treated over two years to take into account the variation in the intensity of treatment from the fact that some teachers receive the treatment for one year and others for two, and some schools have all teachers treated, while others only a fraction.[13] For a school with a single teacher, for example, the variable takes value 1 if the teacher was treated both years, value 0 if she was never treated and

[13] While it is possible that teachers receive treatment for more than two years if they were at a school that had the program for more than two years we are currently unable to follow teachers before 2015.

0.5 if the teacher was treated on only one year. Therefore we expect that the treatment, which provides one year of training for all teachers in the school in 2016, will increase this variable by 0.5 for treated schools.

Columns 4-5 and 9-10 of Table 2 present the 2SLS estimates of the treatment on learning outcomes, instrumenting the proportion of treated teachers using the random treatment assignment. This 2SLS effectively scales up the OLS coefficients presented in Columns 2 and 7. Columns 4 and 9 show the results of the dummy variable considering a school as treated if it received a previously treated teacher or was treated itself. The coefficient on our ITT estimate from column 2 increases slightly adjusting for the fact that some control schools received treated teachers. Columns 5 and 10 show the same 2SLS estimates but now using the proportion of teachers treated over two years. We see that the ITT coefficients double, since treatment increases the proportion of treated teachers by roughly fifty percentage points as expected. Therefore, once we account for variation in the proportion of treated teachers, we find that the program improves standardized test scores by 0.25 SD in reading comprehension and 0.38 SD in mathematics for schools that had all its teachers treated for two years.[14]

4.1.2 Student Skill Distribution

Teachers could react to the treatment in various ways: for example, they could focus their efforts on the lower end of the distribution to bring failing students up, they could focus on top students shifting resources away from those who struggle, or they could acquire skills to strengthen their ability to engage with students throughout the entire skill distribution. In order to test which part of the student grade distribution is shifting in response to the treatment we run a quantile regression taking advantage of the availability of individual student test scores.

Figure 2 shows the coefficient results for quantile regressions at each of the deciles of the test score distribution, including school district fixed effects as in our preferred specification (Columns 2 and 4 of Table 2). We find that the treatment is effective in improving test scores along most of the student distribution and we cannot reject that treatment effects are constant across all deciles. This suggest that the program, which focuses on

[14] Since the program was originally designed to treat schools for 3 years, we are not capturing the full impact of the program as intended. Assuming that the program effects continue to increase we could be underestimating the impact of the full program.

individual teacher weaknesses, is helping teachers to deal with the particular challenges their students face regardless of their position on the ability distribution.

4.2 Local School Outcomes

While the coaching program has large effects on standardized test scores, we also test whether it affects local school outcomes like student grades assigned by teachers or grade repetition that are more visible to students and parents. While grades are not comparable across schools since teachers may have different grading standards, students are supposed to be graded on a 4-point scale that corresponds to their mastery of the standardized curriculum for second grade. In theory, if not in practice, this should reflect roughly the same information as the ECE test scores. However, as we can see from Table A.3 this does not appear to be the case. While only 19% of the students in our sample score “satisfactory” on the ECE (Level 3, which implies mastery of the material), this proportion is almost 90% when measured by classroom grades.

Nevertheless, using classroom grades and grade repetition has the advantage that we can look at impact on a wider set of schools. While only schools with 5 or more students in second grade take the standardized test scores, all schools report classroom grades and grade repetition to the Ministry in this database.

Table 3 shows that the coaching program has no effect on student grades according to the evaluations done by the teachers. Columns 1-8 show the effect of the treatment on school grades for our two preferred specifications: Column 1 and 5 control for school district fixed effects, while Columns 2 and 6 use the panel data and include school and year-by-region fixed effects. Both specifications find an estimated impact of a fairly precise zero on student grades within the school, a finding that contrasts sharply with the fact that students perform much better in the standardized test scores. These results are robust to adjusting for teacher rotation in the 2SLS specification that uses the random assignment as an instrument for the proportion of treated teachers in the school (Columns 3-4 and 7-8 of Table 3).

There are two possible explanations for this finding. First, it could be that teachers grade students using an implicit class curve, comparing students to their peers rather than according to absolute learning standards. This would be consistent with the fact that the entire student distribution shifts to the right, which could mean that the classroom mean simply shifts leaving classroom grades unaffected. An alternative explanation is that tutors are making teachers better at preparing students for the standardized test score, or "teaching to the test." However to the extent that the test is intended to capture the skills that a student in second grade is supposed to master, we may expect that it would affect for example the number of students who fail second grade due to the fact that they failed to master the material.

However, the results in Columns 9-12 of Table 3 show that there is no impact of the program on grade repetition. They show the difference in means in grade repetition between the treatment and control schools, including school district (Column 9) and school fixed effects (Column 10), and the 2SLS results adjusting for teacher movements (Columns 11 and 12). These results suggest that the teachers are grading on a curve, since if they graded on an absolute scale (as they are supposed to) we would have expected that increases in mastery of the material measured by the standardized test scores would yield a lower grade repetition rate for treated schools. Moreover, these results and the nature of how teachers assign class grades to students show that even a program that has such strong impacts on standardized tests scores could have no impact on these variables that are much more visible to students and parents, and thus these programs risk not being perceived to be particularly effective by the direct beneficiaries.

5 Channels

We have thus far shown that the program has a large and significant impact on student learning outcomes after one year of the program, and even larger impact if we adjust for variation in the intensity of treatment (whether teachers receive the program one or two years) generated by teacher movements. We now explore the channels through which the program impacts student learning outcomes.

The program is designed to build up the human capital of teachers. However, it is possible that the tutors who show up to distant schools that otherwise receive little to no attention from the local or national governments could be serving the function of monitors: incentivizing teachers to show up and to exert more effort while the tutor is present in the classroom. Additionally, in remote areas of the country it is possible that the teachers know which week the tutor will show up but not the exact day due to the fact that travel to these schools is arduous and unpredictable. This could motivate them to show up to school every day that week, effectively lowering absenteeism rates in areas where they are notoriously high. If this is the case, we may be better off hiring police (or implementing some other sort of monitoring system) than experienced and expensive coaches.

To distinguish between these two stories of human capital formation versus monitoring we use the experiment of program removal to explore the persistence of the effects after the program ends. We expect that if the treatment is providing teachers with new pedagogical skills, the treatment effect will persist over time as teachers are left with new classroom tools. However, if the program is functioning through monitoring then its effects should disappear once the "monitors" leave.

5.1 Testing Monitoring vs. Skill Development

Table 4 shows the effect of losing the program for those schools randomly assigned to have the program removed, which provides an initial test of the two possible channels of monitoring versus skill development.[15] Columns 1 and 3 of Table 4 show the difference in means between schools that lost and kept the treatment, controlling for school-district fixed effects, while Columns 2 and 4 once again take advantage of the panel data and include school fixed effects. The results show that there is a strong negative impact of losing the program on the outcomes of these schools compared to schools that kept the treatment. This drop is similar in magnitude to the effect of acquiring the program for those schools in the experimental sample that gained the program. In other words, schools that lose the program the very next year appear to lose the entire effect of the treatment.

The fact that the treatment effect entirely disappears one year raises a puzzle, and seems to suggest that the treatment is dependent on the tutors' presence, which would be consistent with the monitoring story. However, we have to remember that the Peruvian system is characterized by high levels of teacher rotation. To what extent are we simply capturing the fact that teachers are moving? It is, after all, the teachers who are being directly treated by the program so the effect could be persistent if we follow teachers.

[15] Figure A.1 in the Appendix shows that the control and treatment groups for this subset of the experimental sample are balanced in pre-treatment outcomes and key covariates, with the exceptions of numbers of students and teachers, which we control for as in the first experimental sample.

5.1.1 Empirical Challenge: Teacher Movement is Endogenous

In order to be able to adjust the effect for teacher rotation, we first have to address the empirical challenge that teacher movements are endogenous, and in particular that they could be either driven by or correlated with our treatment in ways that bias the estimate. We first, therefore, have to ensure two things: on the one hand, that there are no treatment effects of our program on teacher rotation, and on the other, that the treatment does not interact with characteristics that predict teacher rotation.

We can first look at whether the program randomly assigned to schools in 2016 had an effect on teacher rotation in 2017. We are able to track teachers in school databases in January 2017 to see whether the program had an effect on the probability that teachers left the school. For example, on the one hand, one might imagine that treated teachers acquire skills that they can sell on the private school market making them more likely to move, but on the other hand, perhaps teachers with pedagogical support feel more fulfilled at their current school and are less likely to move.

Table 5 shows the treatment effect on subsequent teacher rotation in the subsequent year. We find that, in fact, the program does not affect teacher rotation in 2017 either when measured at the school level (Column 1 shows the proportion of teachers present in 2016 who were still in the school in 2017) or as the probability that any individual teacher stays (Column 2 shows the probability that a teacher (randomly) treated in 2016 stays in his or her same school in 2017 compared to non-treated teachers).[16]

There are no direct effects of the treatment on teacher rotation, but we could still be concerned that there are predictors of teacher rotation that interact with the treatment. In order to identify characteristics of either the teachers, or the sending and receiving schools that are strong predictors of moving we use a Lasso algorithm for model selection. The Lasso is a simple machine learning algorithm that selects the most relevant variables by minimizing the prediction error of the model, while penalizing the coefficients of the regression variables shrinking the coefficients of the variables that are least predictive of teacher rotation to zero.

Table 6 shows the results of the Lasso using a wide selection of teacher and school characteristics. We find that the strongest predictors of moving are being a contract teacher

[16] While there is high attrition in the data since in January many schools have not yet uploaded their teacher rosters, there is no differential attrition for treated and non treated schools in the sample.

(as opposed to having tenure) which is associated with an increase in the probability of moving by 55 percentage points, being younger, and having a higher exam score. At the school level, being in a rural school with deficient infrastructure in terms of ceiling and floor material, water and sanitation make teachers more likely to move. While teachers who move are different from those who do not in important ways, Column 3 of Table 6 shows that these characteristics do not interact with the treatment in any significant way. This is suggestive that teacher rotation is largely orthogonal to the treatment and is unlikely to bias our estimates.

5.1.2 Program Loss Adjusting for Teacher Rotation

We now explore the effects of the program adjusting for teacher rotation. We interact the random removal of the program with the proportion of treated teachers in 2015 that remain in the school in 2016 (as a remainder the second experiment assigns schools treated in 2015 to randomly keep or lose the program in 2016). As can be seen from Table 7 the entire loss of the treatment effect for schools assigned to have the program removed is coming from those schools that no longer kept their trained teachers. In fact, schools that kept 100% of their teachers maintain the program effect performing similarly to those schools in the control group that kept the program.

These results provide strong evidence against the monitoring hypothesis, since it is highly unlikely that the program effect would persist once the tutors stopped coming to the schools if the program was working through a monitoring effect that reduced absenteeism or increased teacher effort while observed. Instead, these results suggest that the program is building up teacher's human capital since schools that retain their teachers also retain the effects of the program even when the tutors are no longer visiting the school.

5.2 Program Persistence when Teachers Move

We test whether the program is affecting only the teacher or if it is having spillovers that affect the entire school environment. We can imagine three different scenarios: First, only the teacher is affected so that when the teacher leaves the entire effect of the program disappears in the original school, but the teacher and her students continue to benefit from the training in new school. Second, there are spillovers on the school (principal, teachers, parents) so that the effects persist even when the treated teacher leaves. A third possibility is that you need both the original treated teacher and the treated school to see the effect of the program.

So far, the results in Table 7 show that once teachers leave the treated school, the full effect of the program disappears. It therefore seems that the program works exclusively through the teacher and is not having spillover effects on the rest of the school (principal, parents, or students) that could persist once the treated teachers leave.

However, do treated teachers retain the effects of the program when they move to non-treated schools, or do you need to have both the treated teacher and the treated school? We could imagine that, for example, if the teacher moves to a new school, he or she would be incapable of translating the training to the new school environment (since perhaps the new school is resistant to these new teaching practices, or the teacher learned very specific tools that are no longer applicable in the new school). In order to answer this question we follow treated teachers who move to non-treated schools and observe the performance of students in the new school.

In order to do this we define two sets of treated schools:

1. *Directly treated schools* are those that received the program in the school (for our experimental sample the identification of the impact is straightforward due to the random assignment of the program)
2. *Indirectly treated schools* are those that received a (randomly) treated teacher who moves there (the movement of teachers is endogenous which raises an empirical challenge).

5.2.1 Difference-in-Differences Identification Strategy

We would like to follow teachers who were randomly given treatment in 2016 who move to non-treated schools in 2017 to test the effect of the program on indirectly treated schools. However, we face the additional challenge that the Ministry discontinued the standardized tests in 2017 due to political pressures and a month-long teacher strike.[17]

[17] Following a very long teacher strike, the Minister felt it would not be appropriate to test students who had not been able to cover the material in the exam.

While the results of the 2018 standardized test scores will be available in early 2019, we are currently unable to use our experimental sample to follow teachers.

As a result, we will rely on an earlier cohort of treated schools and take advantage of the availability of panel data to implement a difference-in-differences estimator over the sample of multigrade, rural schools that were never treated. We estimate the following specification:

$$ECE_{st} = \beta \text{TreatTeacher}_{st} + \theta_s + \alpha_t + \varepsilon_s \tag{3}$$

Where,

- TreatTeacher is a treatment variable that is either a dummy that takes value 1 if the school s received a treated teacher in year t or the proportion of treated teachers in the school.
- θ_s is a school fixed effect, and α_s is year fixed effects. Standard errors are clustered by school.

The identification assumption is that schools that received and did not receive a treated teacher behaved in similar ways prior to receiving the treated teacher suggesting that they would have continued doing so in the absence of treatment. Intuitively we don't want trained teachers to be systematically selecting into schools with different trendlines than alternatives. This is plausible since a majority of the teachers who move are contract teachers who move simply because their contact ends, and many have little choice in where they get sent next. However, the benefit of this empirical strategy is that we can test the parallel trends assumption using data prior to the treatment.

We can also first test the validity of the identification strategy by comparing our experimental estimates of the treatment effect to what this non-experimental approach would yield. Table 8 shows the results comparing the experimental and non-experimental (difference-in-difference) estimates of the treatment effect. We find that the difference-in-difference identification strategy yields very close estimates to the experimental results. This, in addition to the fact that the trends for both the treatment and control groups prior to the treatment are strongly parallel (see Figure 3), suggests that the empirical strategy is not biased in a significant way, which allows us to proceed to estimate the effect of the program on indirectly treated schools.

5.2.2 Treatment Effects on Indirectly Treated Schools

Table 9 shows the effects of having a trained teacher move to a non-treated school on the standardized test scores of the students at the new school using the difference-in-difference estimator described in equation 3.[18] Columns 1 and 2 show the results of coding the treatment as a dummy that takes value 1 if the school received any trained teachers. Columns 3 and 4 use the proportion of teachers in the current school that received treatment in their previous schools. On average schools that receive a treated teacher have 25% of its teachers treated. Figure 4 shows that the parallel trend assumptions appear to hold, which is necessary for the identification strategy to be valid.

We find that there is *full* persistence of the treatment effect one year later in the indirectly treated schools. In other words, treated teachers who move to non-treated schools retain the full effects of the program and the students in the new schools benefit from their increased human capital as much as students in the original school did. This suggests that while there are no spillovers on the treated school once the treated teachers leave, the effects of the program persist even if teachers change schools, which means that the program must be providing teachers with skills flexible enough to adapt to their new contexts.

6 Heterogeneity

We examine heterogeneity of the treatment effect on student learning by initial teacher characteristics like contract type, age, experience and ability.

Table 10 shows that the program has a stronger impact for younger teachers, but that there is no differential impact of the program by experience.[19] However, we cannot rule

[18] We identify 2,500 schools in 2016 that received a teacher treated in a previous year. Once we restrict our sample to the subset of multigrade, rural schools of Peru that had never been treated with the program, we are left with 600 indirectly treated schools and 3,000 controls.

[19] We have teaching experience as a categorical variable that takes four values: “7 to 10 years,” “11 to 14 years”, “15 to 20 years”, and “More than 20 years.” However, this variable is only available for a

out that our age is variable is not capturing experience since once we control for our noisier experience variable the effect for the youth variable becomes insignificant. If this is the case it could be that the program is accelerating a learning process that occurs naturally with experience and perhaps could be focused on younger teachers, but it could also be that younger teachers are more open to learning new pedagogical tools than more experienced teachers who are set in their ways.

There is no differential impact of the program for teachers with tenure over annual contracts, as shown in Columns 3 and 6 of Table 10. Again this is consistent with the fact that the program is working through the pedagogical skills channel, since if it the tutors were serving merely as monitors we would have expected that tenure would mitigate the impact of the treatment due to the job security making teachers less concerned with improving their performance. The fact that the program works as well for tenured teachers as contract teachers also provides a policy alternative to improving teachers that is much more politically feasible than replacing tenured teachers with contract teachers, which would face strong opposition from labor unions.

Table 11 shows the treatment effect interacted with initial teacher quality as measured by the teacher entrance exams.[20] To gain precision, we use the difference-in-difference identification strategy for schools treated between 2011 and 2016 and panel data from 2007 to 2016. The results are similar but slightly underpowered for the experimental sample. Column 1 shows the impact of the program interacted with a treatment dummy that takes value 1 if the average teacher in the school was above the median in the exam scores, pooling teacher entrance exams from 2009, 2010, 2011 and 2015. Column 2 shows the impact of the treatment interacted with a continuous standardized score variable. Both specifications show that the program disproportionately benefits teachers with higher initial ability as measured by this exam. More specifically, Column 2 shows that having 1 standard deviation higher initial ability makes the program effects 0.06 standard deviations larger in Math and 0.04 SD larger in Reading Comprehension compared to the average teacher.

Looking closer at the exam for 2015 for which we have individual subcomponents allows us to unpack these findings to understand whether being 'higher" or "lower ability" in these tests is capturing cognitive abilities, initial pedagogical skills or content knowledge.

subset of the teachers.

[20] This is the exam teachers take to get into civil service career, which tests cognitive skills and content knowledge of the specific area and grade they were applying to teach.

Columns 3-6 of Table 11 show the interaction of the treatment with the 2015 entrance exam and its subcomponents. Columns 3-4 use the overall 2015 score which confirm the findings that the coaching program is differentially more effective for teachers with higher initial ability for Mathematics, with weaker results for Reading Comprehension. Columns 5-6 and 7-8 look at heterogeneous treatment effects using the first two components of the exam that measure general cognitive skills in math and reading comprehension respectively, while Columns 9-10 use the final subcomponent that measures content knowledge in the specific area the teacher is applying to teach. It appears that both general cognitive ability in reading comprehension as well as content knowledge of their subject or grade matter for the effectiveness of the program.

Therefore, we can conclude that the coaching program is complementary to other teacher skills including cognitive ability and content knowledge of the material they are supposed to teach. While the coaching program is only designed to improve teacher's pedagogical practices, to be most effective at improving student learning the coaching benefits from teachers possessing some level of content knowledge to begin with. These results, therefore, show that these kinds of training programs are complementary to other policies that try to improve selection into the teaching profession (policies like increasing salaries or improving teaching conditions that make the profession more competitive with alternatives) as well as policies that improve content knowledge (for example, improving pre-service training). Importantly, even teachers in the bottom of the distribution of cognitive skills benefit from the program, but there are clear complementarities between the coaching program, initial cognitive ability and initial content knowledge that suggest improving these other margins can yield even larger benefits from a teacher training program of this nature. These heterogeneity results also shed light on the mixed findings of the general teacher training literature, which may be explained in part due to variation in complementary teacher skills.

7 Cost-Benefit Analysis

While this program is very effective in raising student learning outcomes, sending tutors to schools in disperse rural areas is a costly endeavor. In order to gauge the cost-effectiveness of the program we need to make some assumptions about the decay of the program over the years. The cost of benefitting 174,000 students in 2016 was roughly US$40 million.

Therefore, if we only consider the effect of one year, the cost per student is 228 dollars, which yields a benefit of 0.13 SD per 100 USD (the standard measure for comparing education programs).

However, once we take into account the persistence of the program, it becomes much more cost-effective. We make assumptions about two sources of program decay: teacher exit from the school system which we calibrated using our administrative teacher records to be between 5-7% per year, and natural decay of program effects as training starts to wear off or becomes obsolete. In order to calculate the cost-benefit, we begin with the original ratio that a program that costs 40 million dollars can benefit 174,000 students.[21] Assuming that there is a 10% decay, for example, which includes teacher exit and decay of the treatment effect, the first year 174,000 students are treated, the next year 156,600, the third 141,000 and so on. Our final assumption is that teachers retire at age 65, and since the average treated teacher is 40 years old, after 25 years the program effects cease altogether. This yields a total amount of students benefited to be between 867,246 and 2.5 million depending on the assumptions.

Table 12 shows that under total decay rates of 5-10% the program becomes much more cost-effective with benefits of up 0.72 to 1.12 SD per 100 USD. We use a simple average of the estimated treatment effects in mathematics and reading comprehension for simplicity. Columns 1-3 incorporate the costs and estimated impacts of one year of training, while Columns 4-6 include the costs of two years of training, as well as the estimated impact which we get from those teachers trained for two years in Table 2. Taking an annual decay rate of 10% we have cost-effectiveness of between approximately 0.70.[22]

8 Conclusion

In this paper, we provide the first evidence of a randomized coaching program for teachers implemented at scale by the government. We find that this kind of program, which provides continuous support to teachers over the entire school year focusing on practical classroom skills, has large, positive effects of between 0.25 and 0.38 SD on student learning outcomes measured in standardized test scores. Moreover, we find that all students along

[21] This assumption could be violated if expanding it becomes more expensive as schools become increasingly more dispersed.

[22] We assume that the second year of treatment has the same cost as the first.

the test score distribution are benefitting from the program.

The program effects are also highly persistent. We find that once we take into account high teacher rotation, the program effects persist in schools that lose the program as long as the trained teachers remain. These results suggest that tutors are building up teacher's human capital rather than serving as monitors. At the same time, the fact that the schools that lose the program and lose their treated teachers experience a large drop in test scores, suggests that the program works exclusively through the teacher and is not having spillover effects on the rest of the school (principal, parents, or students) that could persist once the treated teachers leave. We are able to follow treated teachers to non-treated schools and find that the entire effect of the program persists one year later so that students in the new school benefit from the program as much as students in the original school.

The high mobility of teachers, however, implies that the government has a fundamental role in providing training for teachers since individual schools are likely to underinvest given the likelihood that teachers leave taking their human capital investment with them. This is similar to findings in the labor and public finance literature that firms will underinvest in general worker training since workers can move taking their human capital investment to other firms. This teacher turnover also has implications for government policies that try to decentralize spending to local governments or even schools, or policies that result in significant privatization of school systems, which, if teachers are mobile, could lead to underinvestment in any kind of program that targets the human capital formation of teachers.

Exploring heterogeneous impacts by teacher characteristics, we find that the program works best for younger teachers, and that there is no differential impact by whether teachers have tenure (job-security). Finally, we find that teachers with higher initial cognitive skills and with higher content knowledge of their teaching area benefitted the most from this program. This suggests strong complementarities between the pedagogical training provided by the treatment and other policies intended to improve either teacher selection into the profession or strengthen content knowledge. These heterogeneity results suggest that the mixed findings of the general teacher training literature may be explained in part due to variation in complementary teacher skills.

We calculate the cost-effectiveness of the coaching program, and while it is expensive, once the persistence of the program is taken into account, the program becomes relatively cost-effective. As a result, we can conclude that contrary to the focus of the literature on incentive schemes and monitoring over teacher training, we have strong evidence that teacher training that provides regular support with practical classroom skills can be both effective and persistent at raising student test scores. This is a promising public policy for those countries like Peru that need to upgrade the skill set of their current stock of teachers in order to improve the quality of their public education systems.

References

Albornoz, F., Anauati, M. V., Furman, M., Luzuriaga, M., Podesta, M. E., and Taylor, I. (2018). Training to teach science: experimental evidence from Argentina. *World Bank Economic Review.*

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., and Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333(6045):1034–1037.

Bal, D. and Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In Darling-Hammond, L. and Sykes, G., editors, *Teaching as the learning profession: Handbook of policy and practice*, pages 3–32. Jossey-Bass, San Francisco.

Barrera-Osorio, F., Garcia, S., Rodriguez, C., Sanchez, F., and Arbeláez, M. (2018). Concentrating efforts on low-performing schools: Impact estimates from a quasi-experimental design. *Economics of Education Review*, 66:73–91.

Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy*, 70(5):9–49.

Biancarosa, G., Bryk, A. S., and Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal*, 111(1):7–34.

Bruns, B. and Luque, J. (2014). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* World Bank, Washington, DC.

Campbell, P. and Malkus, N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, 111(3):430–454.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632.

Cilliers, J. and Taylor, S. (2017). Monitoring Teachers and Changing Teaching Practice: Evidence from a Field Experiment. *Working Paper.*

Coalition for Evidence-Based Policy (2013). Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects. http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf.

Duflo, E., Dupas, P., and Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, 123(C):92–110.

Duflo, E., Hanna, R., and Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4):1241–78.

Evans, D. and Popova, A. (2016). What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Review. *World Bank Research Observer*.

Ganimian, A. and Murnane, R. (2016). Improving educational outcomes in developing countries: Lessons from rigorous evaluations. *Review of Educational Research*, 86(3):719–715.

Glewwe, P. and Muralidharan, K. (2015). Improving School Education Outcomes in Developing Countries. *RISE Working Paper*, 15/001.

Hill, H. C., Corey, D. L., and Jacob, R. T. (2018). Dividing by Zero: Exploring Null Results in a Mathematics Professional Development Program. *Teachers College Record*, 120(6).

Kraft, M. A., Blazar, D., and Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*.

Kremer, M., Brannen, C., and Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. *Science*, 340.

Matsumura, L. C., Garnier, H. E., and Spybrook, J. (2012). The effect of content-focused coaching on the quality of classroom text discussions. *Journal of Teacher Education*, 63(3):214–228.

Matsumura, L. C., Garnier, H. E., and Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, 25:35–48.

Ministerio de Educación de Perú (2016). *Manual de Acompañamiento Pedagógico*. Ministerio de Educación de Perú- Area de Fortalecimiento de Capacidades, Lima, Perú.

Muralidharan, K. and Niehaus, P. (2017). Experimentation at Scale. *Journal of Economic Perspectives*, 31(4):103–124.

Muralidharan, K. and Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1):39–77.

Muralidharan, K. and Sundararaman, V. (2013). Contract Teachers: Experimental Evidence from India. *NBER Working Paper No. 19440*.

OECD (2016). *PISA 2015 Results: Excellence and Equity in Education*. OECD Publishing, Paris.

Schleicher, A. (2016). *Teaching Excellence through Professional Learning and Policy Reform: Lessons from around the World*. OECD Publishing, Paris.

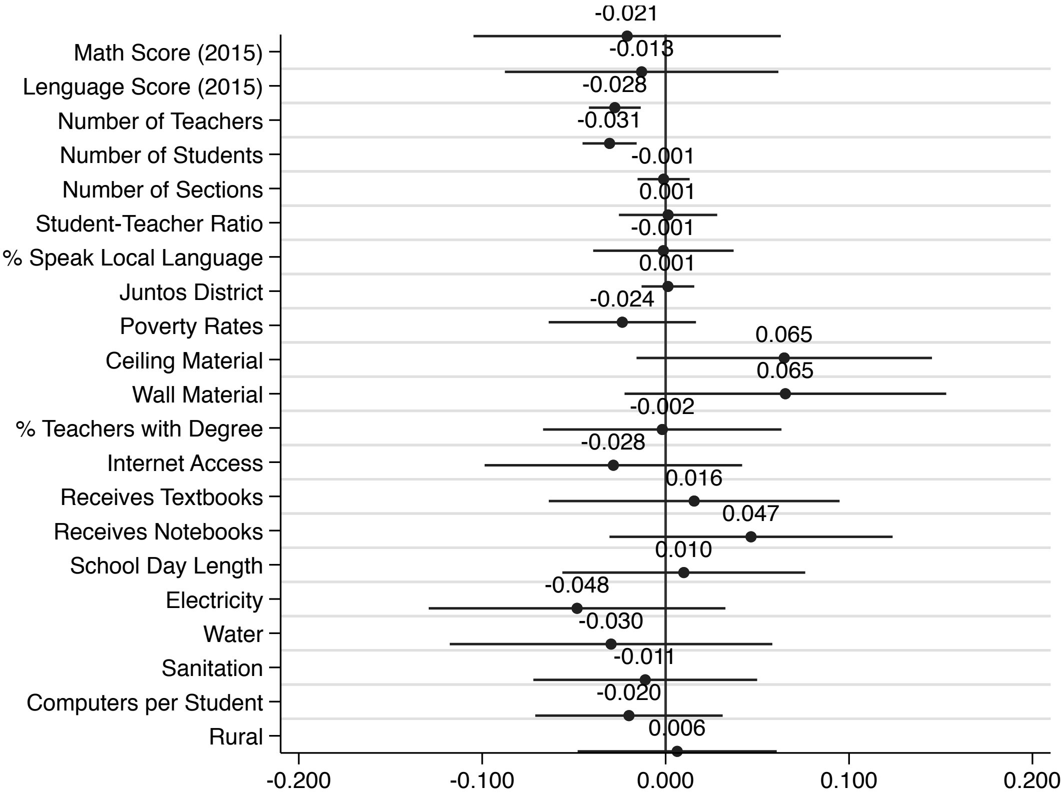
Villegas-Reimers, E. (2003). *Teacher Professional Development: An International Review of the Literature*. UNESCO International Institute for Educational Planning., Paris.

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

Yoon, K., Duncan, T., Lee, S., Scarloss, B., and Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement. *Issues and Answers Report*, 33:3–32.

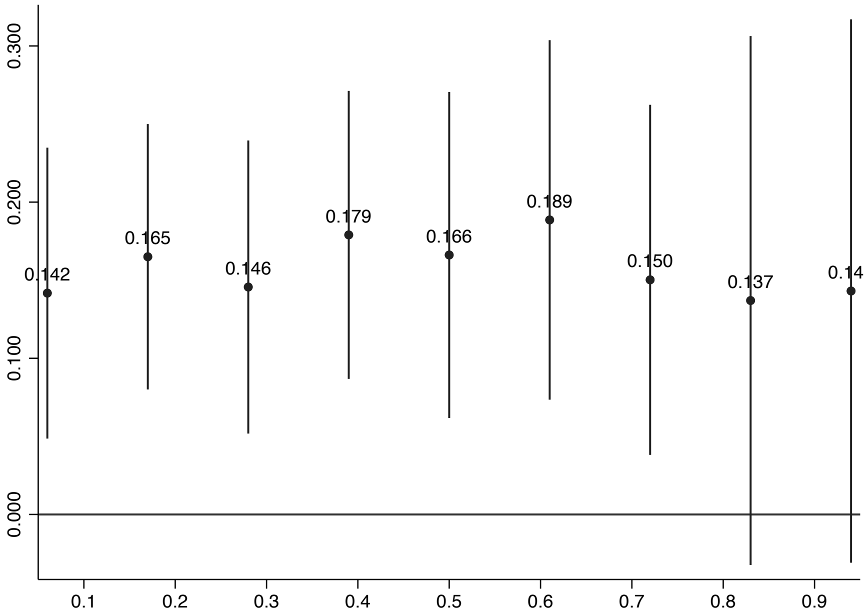
Yoshikawa, H., Leyva, D., Snow, C., Treviño, E., Barata, M. C., Weiland, C., Gomez, C. J., and Moreno, L. (2015). Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes. *Developmental Psychology*, 51(3):309–322.

Figure 1: Baseline Covariate Balance- Standardized

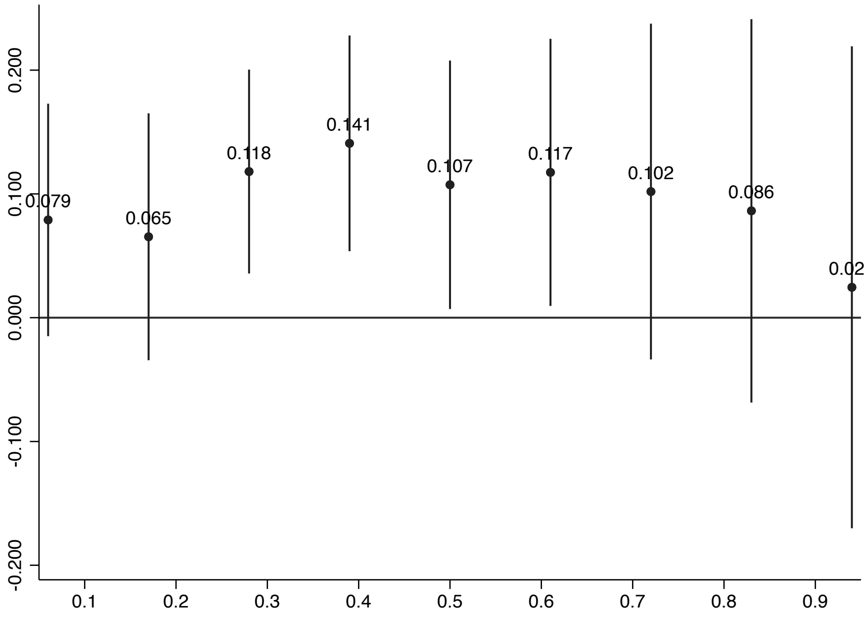


Note: This figure shows the coefficients and 90% confidence intervals for the baseline covariate balance. All regressions include region fixed effects since the program randomization was stratified by region, and is restricted to those schools that have standardized test scores for 2016. All variables are standardized for comparability. This figure shows the entire experimental sample, while Appendix figure A.1 shows them separately for the two experimental groups. Baseline covariates come from administrative databases including NEXUS, SIAGIE, and Censo Escolar.

Figure 2: Quantile Regression Results



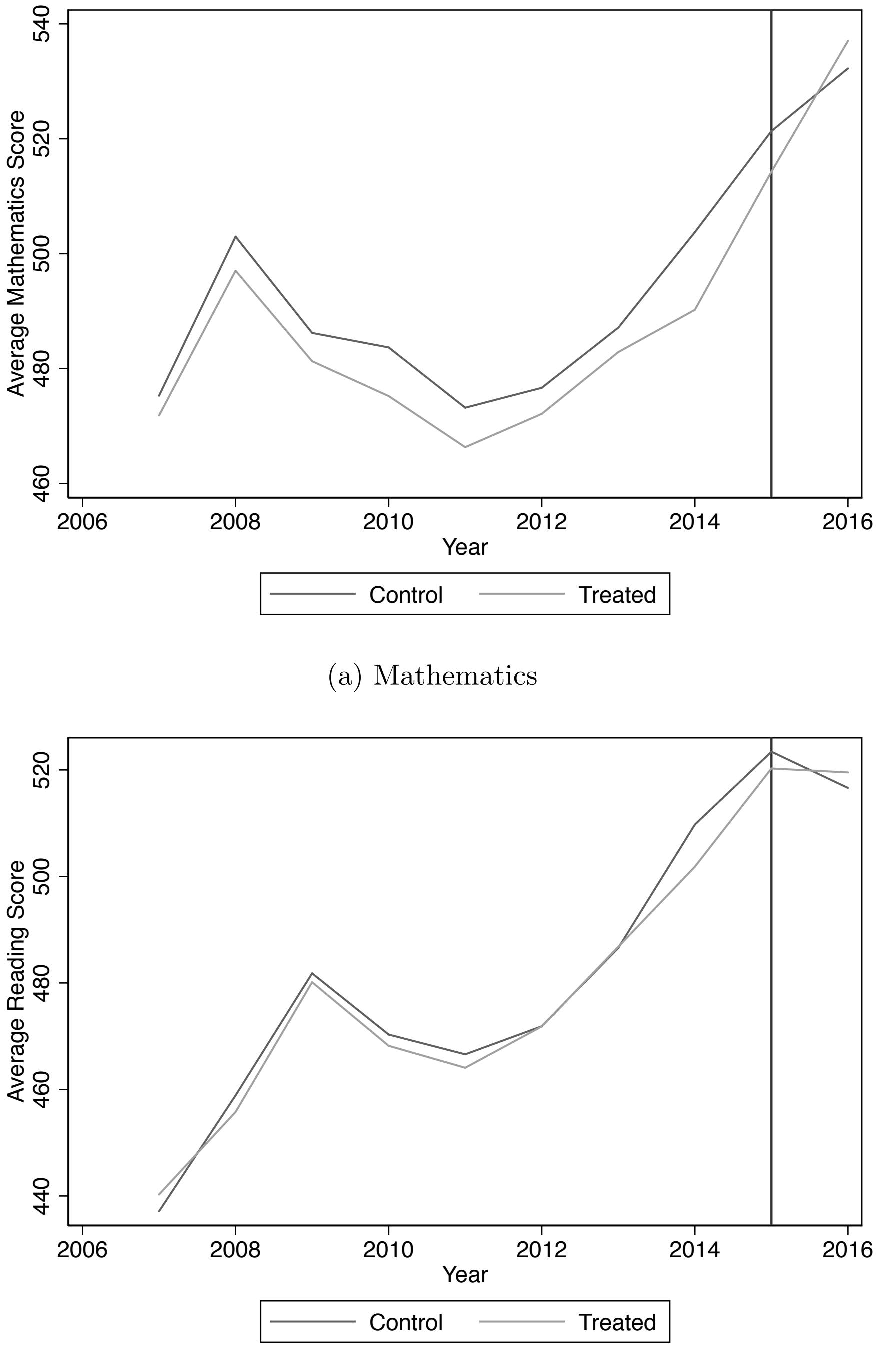
(a) Mathematics



(b) Reading Comprehension

Note: These figures show the quantile regression coefficients for the effect of the program on standardized test scores for each decile of the distribution of student test scores. 95% C.I. shown with standard errors clustered by school. All specifications include school district fixed effects and control for school size.

Figure 3: Common Trends on the Program Effect

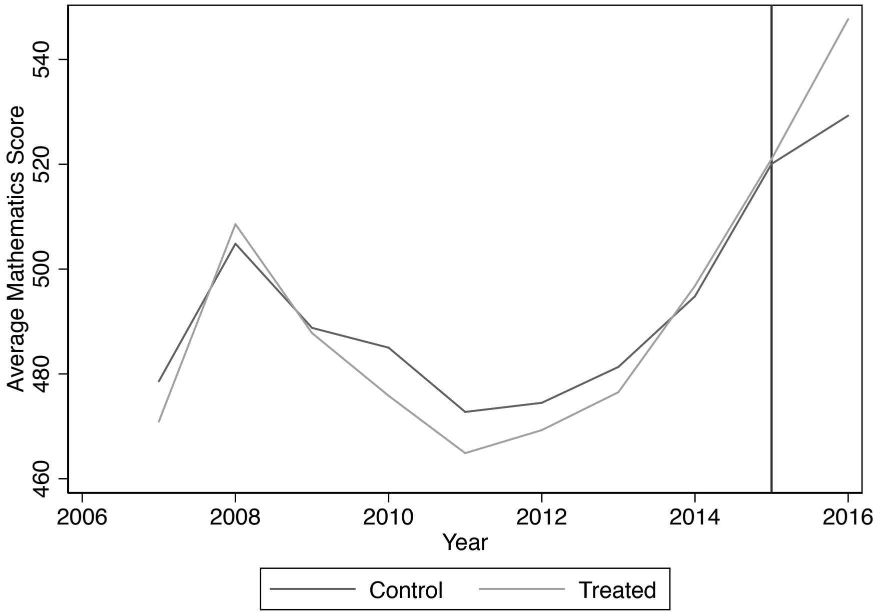


(a) Mathematics

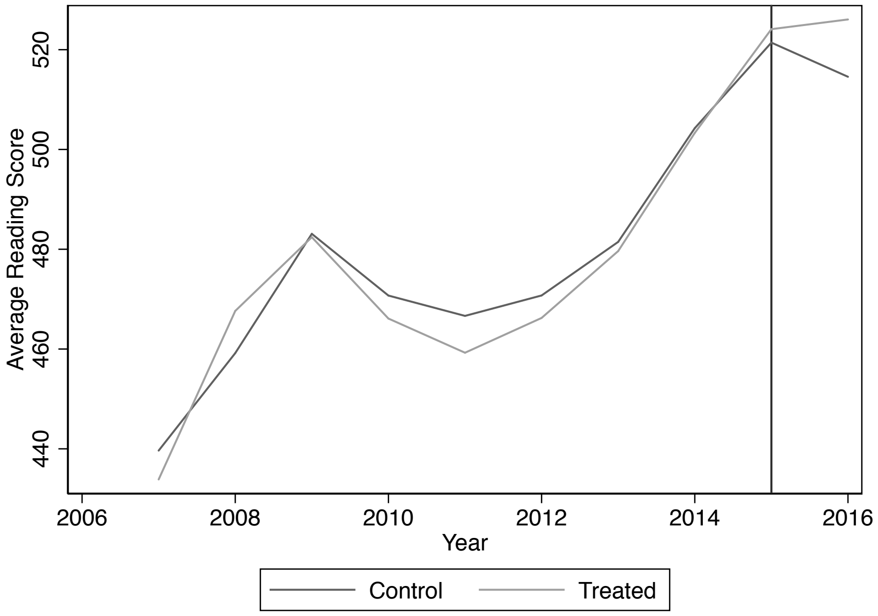
(b) Reading Comprehension

Note: This figure shows the trendline for treatment and control schools for the panel data from 2007 to 2016. Treatment schools are those that received the coaching program in 2016, while control schools are all those schools that are rural, multigrade and did not received the program in 2016. Any school who received the program in previous years has been dropped from the sample.

Figure 4: Common Trends on Indirectly Treated Schools



(a) Mathematics



(b) Reading Comprehension

Note: This figure shows the trendlines for indirectly treated and control schools from 2007 to 2016. The treatment group is defined as those rural, multigrade schools that received a treated teacher in 2016 but had not been directly treated. Control schools are rural, multigrade schools that did not receive a treated teacher. All schools directly treated in any year are dropped from the sample.

Table 1: Descriptive Statistics and Baseline Covariates

	Sample Means		Balance Regressions		
	Control	Treatment	Coefficient	P-Value	N
	(1)	(2)	(3)	(4)	(5)
Math Score (2015)	510.29	513.44	-0.021	0.624	2622
Reading Score (2015)	516.09	520.95	-0.013	0.731	2622
Number of Students	48.37	44.28	-0.031***	0.000	2563
Number of Teachers	2.64	2.51	-0.028***	0.000	2563
Numbers of Sections	5.82	5.84	-0.001	0.879	2563
Student-Teacher Ratio	18.7	18.4	0.001	0.922	2526
Rurality	2.39	2.20	0.006	0.820	2561
% Speak Local Language	4.35	4.22	-0.001	0.951	2563
Poverty Rates (2009)	64.65	56.11	-0.024	0.249	2527
Ceiling Material	5.72	5.81	0.065	0.115	2487
Wall Material	6.07	6.11	0.065	0.144	2487
Floor Material	2.86	2.85	-0.026	0.530	2487
% Teachers with Degree	0.96	0.96	-0.002	0.956	2501
Internet Access	0.10	0.11	-0.028	0.426	1993
Receives Textbooks	0.72	0.76	0.016	0.700	2529
Receives Notebooks	0.68	0.70	0.047	0.236	2528
School Day Length	8.17	8.10	0.010	0.769	2531
Electricity	0.52	0.56	-0.048	0.241	2368
Water	0.54	0.52	-0.030	0.507	2368
Sanitation	0.13	0.15	-0.011	0.721	2368
Computers per Student	0.44	0.46	-0.020	0.443	2376

Note: This table shows the descriptive statistics for the experimental sample, and the regression coefficients for the balance test for the subset of the experimental sample that have standardized test scores available in 2016. The regression coefficients and p-values are shown in Columns 3 and 4. The regressions include region fixed effects since the randomization was stratified by region. Results are identical when including school-district fixed effects. Rurality is a categorical variable that takes values 0 for Urban schools, and 1, 2 and 3 for increasingly rural schools. Ceiling, Wall and Floor Materials are categorical variables that take values up to 7, with higher values implying better materials.

Table 2: Treatment Effect on Students' Standardized Test Scores

	Mathematics					Reading				
	OLS			2SLS		OLS			2SLS	
	Cross-Section		Panel	Cross-Section		Cross-Section		Panel	Cross-Section	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treated	0.156***	0.189***	0.186***	0.216***		0.098**	0.124**	0.120**	0.140***	
	(0.050)	(0.051)	(0.050)	(0.056)		(0.047)	(0.048)	(0.047)	(0.052)	
Prop. Treated					0.379***					0.246***
					(0.098)					(0.092)
FE	Region	District	School	District	District	Region	District	School	District	District
Observations	1852	1852	22938	1852	1838	1852	1852	22933	1852	1838
R^2	0.131	0.230	0.038	0.229	0.228	0.176	0.254	0.061	0.251	0.250
Coefficient on First Stage				0.866	0.494				0.866	0.494
First Stage F-stat				54.42	163.31				54.42	163.31

Note: This table shows the average treatment effect of the coaching program on standardized student test scores. Columns 1 and 6 include region fixed effects, while columns 2 and 7 include school district (UGEL) fixed effects. Both of these specifications control for size of school, which is not balanced at baseline. Columns 3 and 8 include school fixed effects and state-specific time dummies for years 2007-2016, and cluster standard errors by school. Columns 4-5 and 9-10 present 2SLS estimates using the random treatment assignment as an instrument for receiving any treated teachers (Columns 4 and 9) or the proportion of teachers effectively treated in 2016 (Columns 5 and 10). The proportion of treated teachers is coded as the fraction of teachers present in the school in 2016 who were treated in either 2015 or 2016. All results are over standardized exam scores and can be interpreted as standard deviations. Robust standard errors in parentheses, clustered by school in the panel data specifications.

Table 3: Effect on Classroom Grades and Grade Repetition

	School Grades								Grade Repetition			
	Mathematics				Reading							
	OLS		2SLS		OLS		2SLS		OLS		2SLS	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treated	-0.002 (0.008)	-0.000 (0.008)	-0.004 (0.006)		-0.003 (0.008)	-0.000 (0.008)	-0.005 (0.006)		0.000 (0.003)	-0.003 (0.003)	-0.004 (0.004)	
Prop. Treated				-0.005 (0.008)				-0.007 (0.008)				-0.005 (0.006)
FE	District	School	District	District	District	School	District	District	District	School	District	District
Observations	112,695	18,368	112,146	112,136	112,685	18,368	112,685	112,136	114,099	18,368	2,539	2,539
R^2	0.028	0.081	0.029	0.027	0.026	0.081	0.026	0.027	0.015	0.050	0.209	0.210
Coef. First Stage			0.654	0.654			0.498	0.498			0.683	0.495
First Stage F-stat			1,365.18	1364.99			3,677.89	3,677.54			33.33	72.34

Note: This table shows the average treatment effect of the coaching program on classroom grades and grade repetition. Columns 1, 5 and 9 include school district fixed effects, while Columns 2,6 and 10 use panel data and include school and year-by-state fixed effects. Columns 3-4, 7-8 and 11-12 show 2SLS estimates using the random treatment assignment as an instrument for receiving any treated teachers (Columns 3, 7 and 11) or the proportion of teachers effectively treated in 2016 (Columns 4, 8 and 12). Robust standard errors clustered by school in parentheses.

Table 4: Effect of Program Removal

	Math		Reading	
	CS	Panel	CS	Panel
	(1)	(2)	(3)	(4)
Loss	-0.181*	-0.194**	-0.171**	-0.170**
	(0.093)	(0.083)	(0.086)	(0.073)
FE	District	School	District	School
Observations	706	8788	706	8792
R^2	0.332	0.011	0.335	0.033

Note: This table shows the effect of having the program randomly removed for schools that had been receiving it in 2015. Columns 1 and 3 use the cross-sectional data for 2016 and include school district (UGEL) fixed effects and control for school size. Columns 2 and 4 use panel data from 2007-2016 and include school and year fixed effects. All results are over standardized exam scores. Robust standard errors in parentheses.

Table 5: Treatment Effect on Teacher Rotation 2016-2017

	School Level	Teacher Level
	Proportion Stay	Stay
	(1)	(2)
Treated	0.010	-0.008
	(0.027)	(0.023)
Observations	874	811
R^2	0.346	0.063
Mean Dep. Var	0.661	0.895

Note: This table reports the effect of the random assignment of the program on teacher rotation in the subsequent school year. Column 1 uses a measure of teacher rotation at the school level calculated as the proportion of teachers in 2016 present in the school in 2017. Column 2 calculates teacher rotation at the teacher level as the probability that an individual teacher is in the same school in 2017. While there is high attrition for the 2017 data which is reported in January, there is no differential attrition between treatment and control groups. Regressions include region fixed effects and school size.

Table 6: Lasso Predictors of Teacher Rotation

	Lasso Model		Interaction with Treatment			
	Movers	Significance	Coefficient	SE	P-Value	
	(1)	(2)	(3)	(4)	(5)	
Tenure	-0.5492	***	-0.002	0.028	0.941	
Age	-0.0074	***	0.000	0.001	0.778	
Teacher Score (Std)	0.0323	***	0.022	0.015	0.131	
Rurality	0.0261	***	-0.005	0.013	0.728	
Multigrade	0.0270	***	-0.053	0.039	0.177	
Ceiling Material	-0.0009	***	0.002	0.006	0.785	
Floor Material	-0.0028	***	0.007	0.013	0.570	
Receive Notebooks	0.0032	***	0.019	0.022	0.378	
District Poverty	0.0002	***	0.000	0.001	0.377	
Water	-0.0053	***	-0.039	0.022	0.066	*
Sanitation	-0.0032		-0.009	0.028	0.749	
R-squared	0.5282		0.556			
Observations	63,927		5,097			

Note: This table shows the predictors for teacher rotation and whether they interact with the treatment. Columns 1 and 2 of this table show the OLS regression coefficients for the variables selected by the lasso algorithm for model selection using all available teacher and receiving school variables to predict teacher rotation. The dependent variable is a dummy that takes value 1 if the teacher switched schools at the end of the year. Columns 3-5 shows the coefficients for interactions between those same variables and the randomly assigned treatment.

Table 7: Effect of Program Removal Adjusting for Teacher Rotation

	Math (1)	Reading (2)
Loss	-0.365**	-0.435**
	(0.173)	(0.171)
Prop. Treated 2015	0.280*	0.211
	(0.149)	(0.150)
Interact	0.265	0.389*
	(0.245)	(0.232)
Observations	703	703
R^2	0.344	0.347

Note: These regressions show the effect of the randomly assigned removal of treatment when it is interacted with the proportion of treated teachers that remain in 2016. All regressions include school district (UGEL) fixed effects and control for school size. Robust standard errors are shown in parentheses.

Table 8: Comparing Experimental and DD Estimates of the Treatment Effect

	Experimental		Non-Experimental	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Treated	0.189***	0.124***		
	(0.051)	(0.048)		
Treated_{it}			0.191***	0.117***
			(0.049)	(0.045)
Observations	1852	1852	33448	33448
R^2	0.230	0.254	0.013	0.033

Note: This table shows compares the experimental and non-experimental (difference-in-differences) estimates of the randomly assigned treatment effect of the coaching program on standardized test scores in 2016. Columns 1 and 2 replicate the results shown in Table 2. Columns 3 and 4 show the results of the DD estimator which uses panel data from 2007 to 2016 and all non-treated multigrade rural schools as controls. Robust standard errors in parentheses, clustered by school in columns 3 and 4.

Table 9: Treatment Effects on Indirectly Treated Schools

	Dummy		Proportion	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Treated Movers	0.173**	0.164**		
	(0.072)	(0.066)		
Prop. Treated Movers			0.336**	0.301**
			(0.155)	(0.141)
Observations	30706	30706	30706	30706
R^2	0.014	0.033	0.014	0.033

Note: The table shows the effects of having a trained teacher move to a non-treated school on the standardized test scores of the students at the new school using a difference-in-difference estimator on the subset of multigrade, rural schools of Peru. Columns 1 and 2 code a dummy that takes value 1 if the school received any trained teachers, while Columns 3 and 4 use the proportion of teachers in the school that received treatment in their previous schools. Robust standard errors clustered by school in parentheses. All specifications include year school and year fixed effects and include panel data from 2007 to 2016.

Table 10: Heterogeneity by Teacher Experience and Tenure: Panel Data

	Mathematics			Reading Comprehension		
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.127**	0.172***	0.171***	0.059	0.102***	0.107***
	(0.055)	(0.042)	(0.043)	(0.048)	(0.038)	(0.039)
Int. Exper.	-0.001			0.004		
	(0.032)			(0.027)		
Int. Age		-0.049*			-0.067**	
		(0.029)			(0.027)	
Int. Tenure			0.005			-0.012
			(0.027)			(0.025)
Observations	16435	29540	28853	16433	29540	28854
R^2	0.007	0.009	0.009	0.036	0.033	0.032

Note: This table shows heterogeneous treatment effects by experience, age and contract type using the panel data specification for precision. Columns 1 and 4 show the treatment effect interacted with experience, a categorical variable that takes value 1 for 7-10 years of experience, values 2 and 3 for intermediate levels and 4 for more than 20 years of experience. Columns 2 and 5 show interactions of the treatment with the average age of the teachers in the school, while Columns 3 and 6 interact the treatment with the proportion of teachers in the school with tenure. Robust standard errors in parentheses, clustered by school. All specifications include school and year fixed effects. All variables are standardized so that coefficients can be interpreted as standard deviations.

Table 11: Heterogeneity by Teacher Quality: DD Panel Sample

	All Years		2015 Exam							
	Overall Score		Overall Score		C1: Math		C2: Reading		C3: Teaching Area	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Panel A: Mathematics					
Treated	0.178***	0.204***	0.179***	0.210***	0.206***	0.208***	0.176***	0.209***	0.186***	0.210***
	(0.017)	(0.013)	(0.018)	(0.014)	(0.018)	(0.014)	(0.018)	(0.014)	(0.018)	(0.014)
Treat×Dummy	0.072***		0.061**		0.000		0.066**		0.045*	
	(0.024)		(0.026)		(0.026)		(0.026)		(0.026)	
Treat×Score		0.055***		0.031**		0.012		0.038***		0.031**
		(0.013)		(0.015)		(0.014)		(0.014)		(0.015)
Observations	60715	60715	51677	51677	51677	51677	51677	51677	51677	51677
R^2	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
					Panel B: Reading Comprehension					
Treated	0.152***	0.171***	0.154***	0.168***	0.177***	0.167***	0.147***	0.169***	0.158***	0.169***
	(0.016)	(0.012)	(0.017)	(0.013)	(0.017)	(0.013)	(0.017)	(0.013)	(0.017)	(0.013)
Treat×Dummy	0.052**		0.029		-0.022		0.043*		0.021	
	(0.022)		(0.023)		(0.023)		(0.023)		(0.023)	
Treat×Score		0.039***		0.012		-0.004		0.022		0.014
		(0.013)		(0.013)		(0.013)		(0.013)		(0.013)
Observations	60714	60714	51680	51680	51680	51680	51680	51680	51680	51680
R^2	0.030	0.030	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032

Note: This table shows treatment heterogeneity by initial teacher quality as measured by the teacher entrance exams. We use the difference-in-difference identification strategy for schools treated between 2011 and 2016 and panel data from 2007 to 2016. Column 1 shows the impact of the program interacted with a treatment dummy that takes value 1 if the average teacher in the school was above the median in the exam scores, pooling all teacher entrance exams from 2009, 2010, 2011 and 2015. Column 2 shows the impact of the treatment interacted with a continuous standardized score variable. Columns 3-6 repeat these two specification for the 2015 entrance exam and its subcomponents: Columns 3-4 use the overall 2015 rank, Columns 5-6 and 7-8 use the first two components that measure general cognitive skills in math and reading comprehension respectively, while Columns 9-10 use the final subcomponent that measures content knowledge in the specific area the teacher is applying to teach. Panel A shows the results on the outcome variable of students standardized scores in Mathematics, and Panel B for Reading Comprehension. All specifications include school and year fixed effects, and cluster standard errors by school.

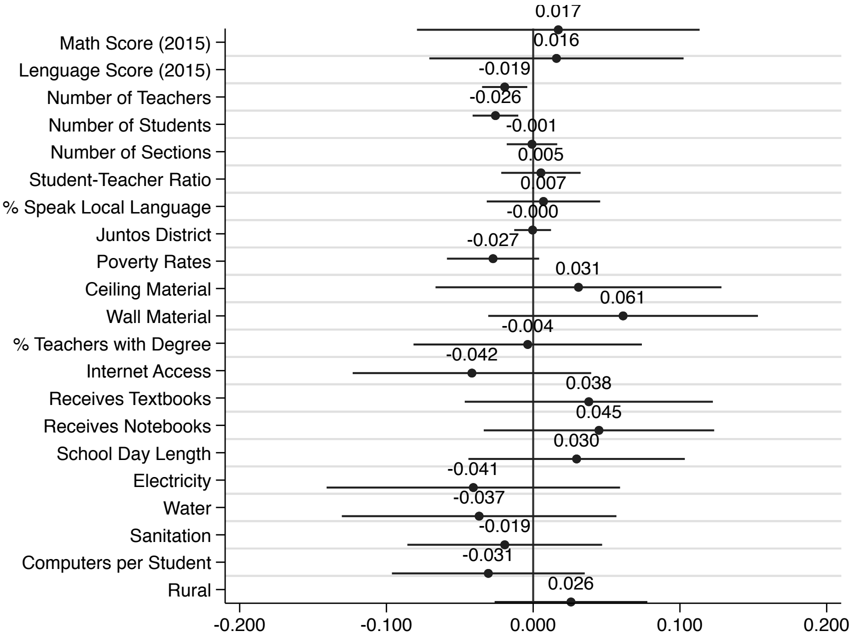
Table 12: Cost-Benefit Analysis Under Various Decay Scenarios

	One Year			Two Years		
Decay	Students Attended	Cost per Student	SD per 100 USD	Students Attended	Cost per Student	SD per 100 USD
	(1)	(2)	(3)	(4)	(5)	(6)
5%	2,516,231	15.9	1.12	2,639,500	30.3	1.04
10%	1,616,079	24.8	0.72	1,776,298	45.0	0.70
15%	1,140,752	35.1	0.51	1,311,335	61.0	0.52
20%	867,246	46.1	0.39	1,040,531	76.9	0.41

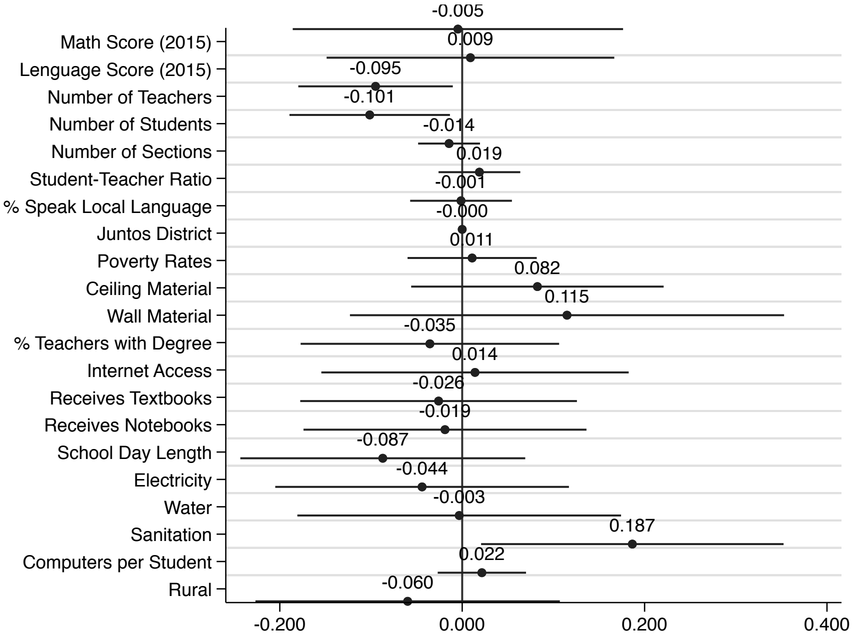
Note: The table shows the cost-effectiveness of the program under various decay scenarios simulated as teachers exiting the school system plus some normal wearing off of program effects as knowledge fades or becomes obsolete. Columns 1-3 show the different scenarios under the assumption that the program lasts one year which assumes the costs and benefits estimated for those teachers that received one year of the program, while Columns 4-6 assume that 2 years of coaching are required for full training effects. We calculate the total number of students that would benefit from the trained teachers under these various scenarios and calculate the cost-effectiveness assuming that the program costs remain constant.

A Supplementary Material

Figure A.1: Baseline Covariate Balance by Experiment



(a) Experiment 1: Random Assignment



(b) Experiment 2: Random Removal

Note: This figure shows the coefficients and 90% confidence intervals for the baseline covariate balance for each of the experimental groups separately: Panel A includes schools in the experimental sample that randomly received the treatment, and Panel B for schools randomly assigned to lose the treatment. All regressions include region fixed effects since the program randomization was stratified by region, and is restricted to those schools that have standardized test scores for 2016. All variables are standardized for comparability. Baseline covariates come from administrative databases including NEXUS, SIAGIE, and Censo Escolar.

Table A.1: Cost of Coaching Programs

		2016			2017	
Type of School	Budget (USD M)	Students (Thousands)	Schools	Budget (USD M)	Students (Thousands)	Schools
Multigrade	39.7	174	6,404	44.1	154	6,150
Bilingual	30.5	137	4,112	42.8	129	3,743
Single-Grade	55.4	585	3,218	40.3	889	3,211
Rural Secondary	5.5	30	267	4.7	254	267
Total	131.1	925	14,001	131.9	1,426	13,371

Note: This table shows the costs and beneficiaries (schools and students) of the four types of coaching programs in Peru, for the years 2016 and 2017.

Table A.2: Control and Treatment Groups by Region

	Full Sample			With Test Scores		
	Control	Treated	Total	Control	Treated	Total
Amazonas	112	68	180	43	30	73
Ancash	97	62	159	25	9	34
Apurimac	5	6	11	1	4	5
Arequipa	70	123	193	28	36	64
Ayacucho	3	79	82	2	9	11
Cajamarca	432	316	748	210	155	365
Cusco	124	210	334	45	59	104
Huancavelica	93	195	288	17	40	57
Huanuco	232	349	581	108	152	260
Ica	5	175	180	1	73	74
Junin	106	35	141	39	13	52
La Libertad	186	186	372	123	121	244
Lambayeque	90	122	212	42	56	98
Lima	114	269	383	31	80	111
Lima Provincias	0	1	1			
Loreto	244	255	499	113	105	218
Madre De Dios	5	99	104	2	26	28
Moquegua	12	72	84	2	6	8
Pasco	99	102	201	36	26	62
Piura	204	236	440	107	125	232
Puno	8	30	38	4	6	10
San Martin	74	560	634	37	299	336
Tacna	2	57	59	1	10	11
Tumbes	18	34	52	5	9	14
Ucayali	77	154	231	34	60	94
Total	2,412	3,795	6,207	1,056	1,509	2,565

Note: This table shows the control and treatment groups by region for the full sample, as well as for the subsample of schools that took the standardized test in 2016 (only schools with more than 5 students in second grade take the standardized test).

Table A.3: Descriptive Statistics for Outcomes - Baseline

Panel A. Standardized Test Scores (ECE)

	All Schools		In Sample		
	Not Sample	In Sample	Control	Treated	Total
Math					
Rausch	555.49	512.15	510.29	513.44	550.18
Level 1	0.38	0.54	0.55	0.53	0.40
Level 2	0.39	0.32	0.31	0.33	0.38
Level 3	0.23	0.14	0.14	0.14	0.22
Reading					
Rausch	573.08	518.96	516.09	520.95	566.45
Level 1	0.09	0.23	0.24	0.23	0.11
Level 2	0.49	0.58	0.58	0.58	0.50
Level 3	0.41	0.19	0.18	0.20	0.39
N	18,774	2,662	1,070	1,552	21,396

Panel B. Local School Outcomes (SIAGIE)

	2013		2014		2015	
	Control	Treated	Control	Treated	Control	Treated
Math						
Average Grade	2.848	2.859	2.899	2.912	2.932	2.952
Beginning	0.093	0.088	0.077	0.070	0.069	0.060
In Progress	0.012	0.014	0.011	0.012	0.012	0.015
Satisfactory	0.847	0.849	0.855	0.855	0.835	0.838
Outstanding	0.047	0.049	0.063	0.063	0.083	0.087
Reading						
Average Grade	2.846	2.856	2.896	2.908	2.930	2.951
Beginning	0.093	0.088	0.077	0.070	0.069	0.059
In Progress	0.013	0.015	0.011	0.013	0.013	0.015
Satisfactory	0.847	0.850	0.851	0.858	0.837	0.840
Outstanding	0.046	0.047	0.061	0.060	0.081	0.085
Repetition Rate	0.09	0.09	0.08	0.07	0.07	0.06

Note: This table shows the descriptive statistics for the standardized test scores (ECE) in Panel A, and for local school outcomes (SIAGIE) in Panel B. ECE scores are reported as both a standardized Rasch measure with mean of 500 and standard deviation of 100, as well as by categorical levels of achievement. SIAGIE test scores are on a scale of 1 to 4, with 3 and 4 signaling satisfactory and outstanding mastery of the material. Students with average grades below 3 must repeat the year.