Can Teaching Be Taught? Improving Teachers' Pedagogical Skills at Scale in Rural Peru[1]

Juan F. Castro	Paul Glewwe	Alexandra Heredia-Mayo	Stephanie Majerowicz	Ricardo Montero
Universidad del Pacifico	University of Minnesota	Universidad del Pacifico	Universidad de los Andes	University of Minnesota

October, 2024

Abstract

We evaluate the impact of a large-scale teacher coaching program in Peru, a context with high teacher turnover, on teachers' pedagogical skills and student learning. Previous studies find that small-scale coaching programs can improve teaching of reading and science in developing countries. However, scaling up can reduce programs' effectiveness, and teacher turnover can erode compliance and cause spillovers onto non-program schools. We develop a framework that defines different treatment effects when teacher turnover is present, and explains which effects can be estimated. We evaluate this teacher coaching program, exploiting random assignment of that program's expansion to 3,797 rural schools in 2016. After two years, teachers assigned to the program increased their aggregate pedagogical skills by 0.20 standard deviations. The program also increased student learning; after one year, Grade 2 students' mathematics and reading scores increased by 0.106 and 0.075 standard deviations (of the distributions of those test scores), respectively. After three years, the cumulative effect increases slightly, to 0.114 and 0.100, respectively. One reason why these impacts are low is that some uncoached teachers moved into treated schools in years 2 and 3. Following our framework, we estimate that the impacts on students of having a "fully" coached teacher for all three years are 0.18 and 0.16 standard deviations for mathematics and reading comprehension, respectively.

Keywords: education, teacher coaching, pedagogical skill, student learning, teacher turnover.
JEL Codes: I21, O15.

[1] We would like to thank seminar participants at the Department of Applied Economics of the University of Minnesota, the Department of Economics of Universidad del Rosario, the LACEA 2019 Annual Meeting, the Department of Agricultural and Consumer Economics at the University of Illinois, and the Department of Agricultural Economics and Rural Development at Seoul National University for their valuable comments. We also thank several anonymous referees, whose comments were very helpful for improving our paper. We are also grateful to Hugo Fernández for excellent research assistance, and to Diana Horvath for preparing the replication materials. Any remaining errors are ours alone. The randomized evaluation was planned by Peru's Ministry of Education**.** The student assessments and teacher observation instrument used in this study were designed by the Ministry of Education for general internal use. We used anonymized data provided by the Ministry of Education. The replication package for this paper is at: https://doi.org/10.5281/zenodo.13738582.

1. Introduction

Teacher quality is an essential determinant of student learning (Das et al. 2007, Clotfelter et al. 2010, Chetty et al. 2014). Yet many teachers lack mastery in the subjects they teach, or lack the pedagogical skills to teach them effectively. This is especially true for teachers in developing countries (World Bank, 2018). Can these teachers' skills be improved?

Every year, developing countries spend over $1 billion on teacher training (Loyalka et al., 2019). Popova et al. (2016) find that about two thirds of the World Bank educational projects between 2000 and 2012 included in-service teacher training. Such training is attractive because it can be centrally designed and coordinated by the Ministry of Education and is usually supported by teachers' unions (Evans and Popova, 2016).

In this study, we evaluate the impact of a large-scale teacher coaching program, operating in a context of high teacher turnover, on teachers' pedagogical skills as well as on student learning outcomes. Evidence on the impacts of in-service training in developing countries is mixed, and programs vary widely in form and content. A survey by Evans and Popova (2016) found that programs with face-to-face training, follow-up visits, engagement of teachers to obtain their ideas, and adaption to local context, tend to have larger effects on student learning. Coaching programs often have these features as they involve school visits, classroom observations, and personalized feedback for teachers by trained peers or coaches. Thus, coaching programs are a promising alternative to traditional in-service training that offers intensive sessions to large numbers of teachers at a centralized venue.

When programs are offered at the school level but are intended to operate through teachers, and teachers can move between schools, estimates of the average treatment effect (ATE) of the program based on a randomized control trial may be biased. In particular, movement of teachers across schools may lead to spillovers that will introduce biases when comparing treated and control schools, even when all schools comply with their random assignment and there are no biases due to the selection or attrition of students.

Education interventions that operate through teachers often have all teachers in a school share treatment status (i.e., all teachers are either treated or untreated). Most studies of the effectiveness of these types of interventions focus on student outcomes and compare treatment schools with control schools, and some of them evaluate results after enough time has passed for teachers to switch schools (Lucas et. al. 2014, Jukes et. al. 2017, Cilliers et. al. 2020). These studies usually address potential biases due to student attrition, yet they rarely mention the possibility of teacher turnover or the potential bias it may induce.

This risk of bias may occur not only for education interventions but also for any estimation of treatment effects in cluster randomized control trials (RCTs) with movement of service providers or program beneficiaries across clusters. Indeed, high turnover is reported for many non-education contexts. For example, Kovner et al. (2014) report that 17.5% of new nurses in the U.S. leave their jobs within one year of starting, and Banerjee et al. (2021) find, in their control sample, that one-third of police officers in India changed stations over an 18-month period. Despite its frequency, turnover is usually ignored in program evaluations. For example, Georgiadis and Pitellis (2016) compare treated and control enterprises (clusters) in a job training program but do not discuss the possibility of workers moving across firms.

We make a methodological contribution by developing a framework that clarifies the assumptions and data needed to obtain unbiased estimates of average treatment effects (ATE), intent to treat effects (ITT), and average causal response (ACR, an extension of local average treatment effects (LATE)) in a clustered RCT with movement of service providers across clusters. In our context, this framework explains how treatment effects differ, depending on whether one focuses on a particular set of teachers, following them if they move to other schools (in which case the outcome variables are those teachers' skills), or on the teachers and students in particular schools (in which case the outcome variables are the skills of these schools' teachers and the learning progress of these schools' students). Both sets of treatment effects are highly relevant from a policy perspective. The first set is relevant for policies that focuses on improving the skills of a particular group of teachers, such as teachers whose pedagogical skills are thought to be deficient. The second set is relevant for policies aimed at improving the teaching skills and learning progress, respectively, of the teachers and students in a particular group of schools, such as schools where students' academic performance is particularly low.We show how the latter set of effects depends not only on the direct effect of the program on participating teachers' skills but also on the indirect effect of the program on teacher composition: which teachers stay in these schools, which teachers leave these schools, and which teachers move into these schools. Previous research based on cluster RCTs where service providers move across clusters has ignored these composition effects.

We show that, in general, it is not possible to estimate average treatment effects (ATEs) for teacher skill and student learning, although under certain conditions lower bounds for ATEs can be estimated. We also show that comparisons of teachers in treated and control schools after turnover has occurred will, in general, lead to biased estimates of intent to treat (ITT) effects for teachers *in the program schools when the program started*. However, it is possible to estimate these ITT effects if one has a sample of teachers that follows them when

they change schools, or using the data of teachers in treated and control schools after turnover has occurred *if* that turnover is unrelated to the program. This last result is important because following teachers who change schools and, more generally, following service providers who change clusters, can be difficult, which raises the risk of attrition bias in ITT estimates.

We estimate the effects on teachers' pedagogical skills and on student learning of a teacher coaching program implemented in rural multigrade schools in Peru. Trained coaches visit classrooms and give specific advice to teachers on their pedagogical practices, providing customized strategies to improve them. Identification exploits random assignment of 6,218 schools (3,797 treated schools, 2,421 control schools) when the program expanded in 2016. Teacher skills were measured in late 2017 (after nearly two years of treatment) by observing teacher-student interactions and a broad range of instructional practices in a randomly selected subsample of 166 treated and 174 control schools. Student skills were tested in grades 2 (late 2016) and 4 (late 2018) for all public schools with five or more students in those grades, which provides student test score data for 2,567 of the 6,218 randomly assigned schools.

As in many developing countries, Peru's rural schools have very high rates of teacher turnover;[2] of the teachers in the subsample of 340 schools with teacher skills data, about 43% had moved between 2016 and the start of 2017. Importantly, classroom observation data were collected not only in these 340 schools, but also in many (but not all) of the schools that received the teachers who moved from these schools to other schools between 2016 and 2017.

Our main findings are as follows. For the teachers who, after turnover occurred (i.e. in 2017), were teaching in the schools assigned to the program, we find that the ITT effect of two years of coaching on their pedagogical skills is 0.20 standard deviations (s.d.) of the distribution of those skills. This is also our preferred estimate of the ITT effect on the skills of the teachers in the program schools when the program began, some of whom left those schools in the next year. We also show that this ITT estimate is, under plausible assumptions, a lower bound of the ATEs for both sets of teachers. Turning to specific skills, the largest ITT effects are for lesson planning and, to a lesser extent, encouraging students' critical thinking.

We also estimated treatment effects of the program on student learning after one and three years (we have no data for the second year). After one year, the program increased learning among the Grade 2 students who took the 2016 National Student Evaluation by 0.106 s.d. in mathematics and 0.075 s.d. in reading comprehension (of the distributions of those test scores). These are both ITT and ATE effects, since all teachers followed their

[2] High teacher turnover is common in developing countries: Zeitlin (2021) reports turnover of about 20% per year in Rwanda, and Schaffner, Glewwe and Sharma (2024) report 18-21% turnover per year for teachers in Nepal.

random assignment in the first year. After three years of exposure, the ITT effect increases only slightly, to 0.114 s.d. for mathematics and 0.100 s.d. for reading comprehension; these estimates, which are lower bounds for ATE (which cannot be estimated in year 3), reflect the fact that many teachers in program schools in year 3 did not have three full years of coaching, and some teachers who had moved to control schools by year 3 had been coached in previous years. The average causal response (ACR) estimates after three years, which adjust the ITT estimates to estimate the impact of three years of exposure to teachers who were coached in all three years, are 0.180 s.d. for mathematics and 0.162 s.d. for reading comprehension.

Our estimates for the effect of coaching on pedagogical skills are smaller than those found in developed countries (0.49 s.d. on instructional practices, see Kraft et al., 2018). This may reflect the scale of the program, and Peru's high rate of teacher turnover. Yet we address two unresolved questions on coaching's impact on teachers' pedagogical skills in developing countries. We show that: (i) A program implemented at scale, even with high teacher turn-over, can still exhibit positive impacts; and (ii) *General* pedagogical skills can be increased.

Furthermore, while our estimated effects on student learning may seem small, they are similar, and in one sense larger, than those typically found in developing countries. Evans and Yuan (2022) surveyed 224 education studies and found that the median effect on learning outcomes is 0.10 s.d., and these effect sizes decrease with the size of the study. For large studies, those with over 5,000 students, the median effect is only 0.05 s.d.

To our knowledge, no prior study has evaluated the effects on pedagogy and student learning of a large-scale teacher coaching program in a developing country. Most in-service training programs evaluated in those countries are small-scale pilots or efficacy trials run by researchers or NGOs (Evans and Popova, 2016). For example, Cilliers et al. (2020) estimated the impact of coaching and centralized teacher training on student reading skills implemented in 180 public schools in South Africa, and Albornoz et al. (2020) estimated the impact of teacher coaching to improve student learning of science implemented in 70 public schools in Argentina. In contrast, we evaluate a program implemented in 3,797 rural schools in Peru.

The issue of scale is relevant for coaching programs' effectiveness because of two features of this type of in-service training. First, the program's success depends on the supply of qualified coaches. If these skills are scarce, expanding the program likely will reduce its quality, and thus its effectiveness. Second, classroom observation and personalized feedback requires coaches to travel to several schools. This can be costly and can complicate program delivery if scaling-up implies serving schools in very remote areas. This is very likely for rural schools in developing countries, whose teachers often require additional training.

Teacher turnover not only complicates identification of program effects, as discussed above, but may also make coaching less effective by reducing compliance. Teachers who leave a school before the program ends may not receive the full "dose" of coaching, and program schools that receive new teachers may have staff who are only partially coached.

We know of only one other study that considered teacher turnover when evaluating a teacher training program. Matsumura et al. (2010) estimated the effect of a literacy coaching program in 32 elementary schools in Texas. Stressing how such turnover can thwart schools' efforts to improve instruction through teacher training, the authors estimated the program's effect on the reading skills of the students of teachers recruited to replace those who left their school in the first year of the program. They found a positive association between teachers' program participation and their students' reading skills. However, the non-random composition of their sample (recruited teachers in program and non-program schools may not be comparable) casts doubt on the causal interpretation of their results.

Finally, the literature thus far does not provide a clear indication as to whether coaching can improve *general* pedagogical skills. Most evaluations of coaching programs focus on pedagogy for a specific topic or course. For example, Albornoz et al. (2020) focus on improving teaching of science, and Cilliers et al. (2020) focus on reading. Kraft et al. (2018) highlight a lack of causal evidence on the effect of coaching for subjects other than reading or literacy. Some papers measure the effect of training on teacher time allocation (Bruns et al. 2018) or on using specific types of teaching (Kotze et al. 2019), but not on their teaching skills. The pedagogical skills of public-school teachers in developing countries are generally low, and a key policy question is whether coaching can improve a broad set of teaching skills.

The rest of the paper is organized as follows. Section 2 describes the program and explains the evaluation design. Section 3 presents our analytical framework, defines several treatment effects, and explains which can be estimated. Sections 4 and 5 present estimates of the program's impact on teachers' pedagogical skills and on student learning, respectively. Section 6 provides concluding remarks, policy implications, and advice for future research. The Supplemental Appendix (Castro et al. 2024a) contains additional tables and derivations.

2. The Coaching Program and its Evaluation Design

2.1 Teacher Hiring and Movement in Peru

There are two types of teachers in the Peruvian school system: Tenured (civil servant) teachers (*nombrados)*, who have a permanent position in a particular school, and contract teachers (*contratados*) on temporary one-year contracts who are filling in for tenured teachers who are temporarily absent or for unfilled vacancies in particular schools. In the schools we consider – multigrade and monolingual – most (70-75%) of teachers are tenured.

Teachers become tenured through a selection process with two stages. The first stage consists of a nationally administered exam that covers reading comprehension, logical reasoning, and knowledge of pedagogical practices. Teachers with the minimum passing grade on the exam proceed to a second stage that is carried out by regional education offices and includes an interview and in-classroom observation of teaching practices.

Teachers who do not reach a minimum passing grade in exam in the first stage of the selection process, or who receive a passing grade but are unsuccessful at the second stage, can fill temporary teaching positions as contract teachers (and can continue trying to obtain tenure). Contract teachers have annual contracts: at the end of each school year they must apply for either a renewed contract at their current school or for a contract position at another school. When applying to new schools, contract teachers can apply to as many schools as they want within one region. They are then ranked according to their scores on the latest exam, and teachers with the best scores get their top priority of schools. Teachers can maximize their probability of getting placed by ranking as many schools as they are willing to go to, and by selecting less popular schools (for example, schools located in remote rural areas).

Tenured teachers tend to move less frequently given their permanent position in their schools, but they can request a transfer to another tenured position. In order to do this, they must meet three requirements: have been in a tenured position for at least three years, have been in the *current* tenured position for at least 2 years, and cannot move to another school within the same school district (Peru has about 250 school districts (UGELs)).

2.2 The Coaching Program

In 2010, the Peruvian government initiated coaching programs to improve public primary school teachers' pedagogical practices. As per Ministry of Education guidelines, the school district authority (UGEL) hires coaches for teachers in the schools targeted by the program, who are selected from top-performing teachers. Coaches must have a pedagogical college or

university degree, five or more years of primary school teaching experience, and at least one year of experience training or providing support to teachers. Administrative data show that coaches rank much higher than other teachers in the Ministry's teacher evaluations. Coaches were paid the equivalent of US$ 1,200 per month, about double the average teacher's wage.

The Ministry of Education sets the standards for hiring coaches, and for the general program design, but the UGELs select and hire the coaches. Each coach works with eight teachers, and UGELs decide how to match coaches to teachers. Coaches are hired annually. About 20% continue for another year, but only 5% stay in the same school the next year.

The coaching program is a substantial investment by Peru's government, costing over US$ 130 million per year.[3] By 2016, teachers in over 14,000 public schools with more than 900,000 students were being coached under several coaching programs. Over 90% of these schools are primary schools. There are three versions of the program for primary schools: (i) bilingual coaching (for schools where most students speak a Peruvian indigenous language); (ii) monolingual multigrade coaching (for schools where most students speak Spanish and there are fewer teachers than grades taught); and (iii) monolingual full-teacher coaching (for schools large enough to have one teacher per grade and where most students speak Spanish).

This paper evaluates the second type of coaching program,[4] which operates primarily in rural areas.[5] Over 90% of Peru's rural public primary schools are multigrade, which typically have two teachers and about 30 students. Rural multigrade schools are the majority of schools with coaching programs. The monolingual multigrade program is particularly expensive because the target schools tend to be very far apart, so the program requires a large number of coaches and significant travel expenses. This version of the program alone, called *Acompañamiento Pedagógico Multigrado* (APM) in Spanish, cost the government about US$ 40 million in 2016 and served 174,000 students. This implies an annual cost of US$ 228 per student, which is over 20% of the total expenditure per student in Peru's primary schools (in 2015, average spending per primary school student was 2,800 soles, or about US$ 940).

A coach's work consists of several tasks. First, the coach meets the school principal and gathers information about the educational context. Then, the coach attends all teachers' class sessions (one teacher per day) to observe their classroom performance and make an initial diagnostic assessment. The coach uses this assessment to identify the competencies

[3] It was not implemented in 2021 and 2022 due to Covid-19, after which it was restarted, but on a smaller scale.
[4] Although the three types of coaching programs have some differences (such as the teacher-to-coach ratio or the bilingual certification of coaches), what happens during the coaching sessions is very similar in all three types.
[5] About 95% of the 6,218 schools in our study are located in rural areas.

that the teachers must improve and develops an improvement plan with each teacher. During the school year, the coach observes eight more of each teacher's class sessions at regular intervals. The program is usually implemented for three consecutive years. After each classroom observation, the coach and the teacher meet to discuss the progress made in terms of the improvement plan. The coach sends monthly and quarterly reports to the UGEL, and to the school principal, on each teacher's progress and on areas for improvement. At the end of the year, the coach provides a final feedback session for each teacher, collecting his or her impressions of the process, and then writes a final report for each teacher on the achievements, actions, and areas requiring further effort, referencing the initial improvement plan.

In addition to the classroom observations, each coach organizes eight workshops per year for his or her teachers to discuss pedagogical practices and encourage the exchange of ideas. In the workshops, all the teachers for a given coach gather with the coach to discuss a particular pedagogical topic of interest. The coach encourages and guides the exchange of ideas and successful practices among teachers and provides theoretical support on the chosen subject. At the end of each workshop, the group chooses a new topic for the next gathering.

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

> "The pedagogical coaching promotes the development and strengthening of skills related to understanding the student in her context, curricular planning, guiding learning, ensuring a safe school environment, and evaluating student learning. In addition, it promotes the development of critical thinking skills like self-reflection and analysis, through exercises that seek reflection and critical analysis of the teacher's own performance." (APM Manual)

APM uses a cascade system. Each coach is trained, supported and monitored by a pedagogical specialist. Each specialist is required to monitor each coach at least twice per year during the coach's classroom visits. The specialist also provides two workshops per year directly to teachers. Coaches and specialists follow the "Framework for Good Teaching Performance" developed by the Ministry of Education to guide their training. The framework specifies nine skills that teachers should master; the program focuses on seven of these skills:

- Knowledge and comprehension of the students' characteristics and backgrounds.
- Collaborative class preparation with other teaches in the same school.
- Fostering an environment that promotes learning, democratic values, and diversity.

- Guiding the learning process through mastery of the curricular content and the use of effective pedagogical strategies and resources.
- Permanent evaluation of the learning process and provision of feedback to students.
- Active participation in the school management.
- Fostering relationships of respect and collaboration with school community members.

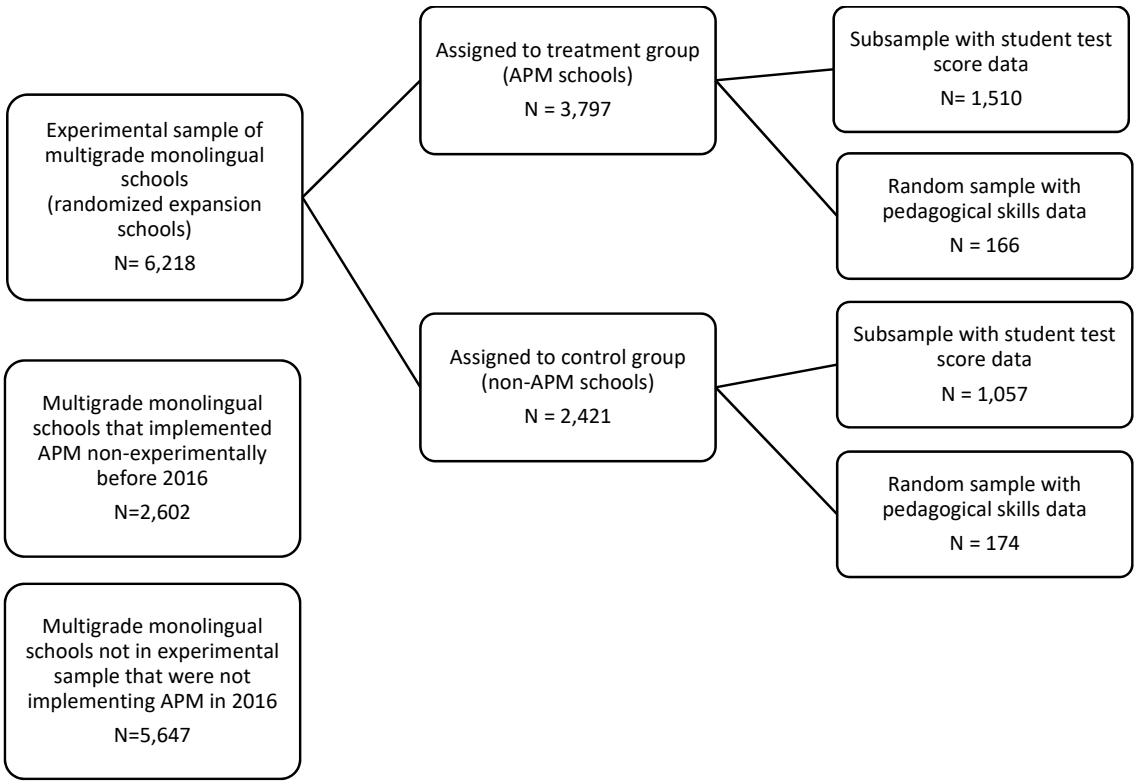
2.3 Evaluation Design

In 2016, the APM program was expanded in a way that involved random assignment. All schools that started APM before 2016, and had not yet completed the full three years of the program, continued to participate in APM and were not part of the experimental sample. Monolingual multigrade schools that had low scores on Peru's Grade 2 national student evaluation and had not yet participated in APM were randomized into treatment and control groups. Of the 6,218 eligible schools, which we call the randomized expansion schools, 3,797 were randomly assigned to the treatment group and started the APM program in February of 2016 (Peru's school year runs from March to November). Henceforth, we call these schools APM schools. The other 2,421 schools, the control group, which we call non-APM schools, did not participate in any coaching program in 2016, 2017 and 2018. This randomization, shown in Figure 1, was stratified at the region (department) level, Peru's highest level of political division (Peru has 26 regions).[6]

The sample size is reduced by the availability of outcome data. Standardized tests scores are available only for the 2,567 schools with five or more students in the grade being tested, and the pedagogical skills were measured only for a stratified (at the region level) random subsample (340 schools, 166 APM, 174 non-APM) of the full experimental sample.[7] Table A1 in Castro et al. (2024a) provides summary statistics for the 6,218 randomized expansion schools (Column (3)), as well as for the wider population of all Peruvian public primary schools (Column (1)), and all monolingual multigrade public primary schools (Column (2)), which is the target population of the APM program. Understandably, the randomized expansion schools tend to be smaller and more rural than the average public primary school, which includes large schools in urban areas. They also differ in access to the

[6] Some regions did not have enough eligible schools to provide equal numbers of APM and non-APM schools, which generated variation in the proportion of APM and non-APM schools across regions. Also, in two regions the number of eligible schools was less than or equal to the quota that they had to fill, which left them without any control schools. Since there was no random assignment in these regions, we exclude them from the analysis.
[7] The initial plan drew a random subsample of 364 schools (182 APM and 182 non-APM), but 24 of these schools could not be reached due to their very remote location.

Figure 1. Samples from Random Assignment to APM and non-APM Schools



internet and to computers, and in the quality of school infrastructure. Differences are much smaller in access to textbooks and workbooks: more than 70% of schools across our samples receive textbooks, and just above 65% receive workbooks. Finally, all schools have a similarly high percentage of teachers with degrees (97%) and similar teacher-student ratios and school-day lengths.

Our sample (Column (3)) is much more similar to the average Peruvian monolingual multigrade school (the target population, Column (2)), with a similar average size of about 29 students and 2 teachers per school, similar access to the internet (8%) and computers, and quality of school infrastructure. We conclude that our sample is very similar to the target population of the APM program, which are monolingual (Spanish-speaking) multigrade schools, but somewhat different from urban primary schools. This should be kept in mind when considering the external validity of our study.

Column (4) of Table A1 shows descriptive statistics for the subset of schools that have test scores. Only schools with five or more students in the tested grade level are eligible to take the national evaluation test, so this subsample includes slightly larger schools than our

full randomized expansion sample in Column (3); the average school has 46 students and 2.6 teachers. Teacher-student ratios are slightly lower in the test score subsample, and several infrastructure quality indicators display small differences, suggesting that subsample schools are slightly better off than, but generally similar to, the full sample. On the other hand, the randomized expansion sample has lower baseline test scores than the average monolingual multigrade school, which reflects that the expansion was targeted towards lower performing schools. Overall, while there are small differences between our test-score subsample and both the full sample and the average monolingual multigrade school, the differences may be small enough to allow one to extrapolate from our sample to all monolingual multigrade schools. Lastly, Column (5) shows means for the subsample of 340 schools where teachers' pedagogical practices were measured. They closely resemble the overall randomized expansion sample (Column (3)), as expected since they are a random sample drawn from those schools.

Timeline. The random assignment was done in late 2015. APM schools began the program in early 2016 and operated it for three consecutive years. The school year begins in March, and the standardized tests are taken in November. We look at effects on students' 2016 and 2018 test scores, one year and three years after the program started. Unfortunately, standardized tests were not administered in 2017 due to a national strike. Our measurement of pedagogical skills took place near the end of 2017, two years after the program's implementation.

Outcomes. The measure of student learning outcomes is the National Student Evaluation (henceforth, ECE, its Spanish acronym) primary school exam that assesses students' mathematics and reading comprehension skills. It has been implemented annually since 2007 and is comparable across years.[8] All schools with five or more students in the tested grade take the exam; this means that some schools move in and out of the testing sample over time. Initially, the ECE tested students at the end of the second grade of primary school but, starting in 2018, it was shifted to fourth grade. This implies that, for our cohort of students, we have test score data at the student level first in 2016 when they were in second grade, and again in 2018 in fourth grade. Table A2 in Castro et al. (2024a) shows descriptive statistics of the exam. The ECE scores are reported both as levels of subject mastery and as a Rasch score with a nationally standardized mean of 500 and standard deviation of 100. Table A2 shows that

[8] The standardized exam was continuously implemented from 2007 until 2016; it was discontinued in 2017 for one year due to a Ministerial decision in that year in response to a prolonged nation-wide teachers' strike. Students had missed several weeks of class and allegedly were not up to date with the subjects that the ECE covered, prompting the decision to cancel it. It was reinstated in 2018. The methodology for the ECE did not change after the gap and the 2018 results are comparable to the ECE exams taken before 2017.

average subject mastery is low; a large proportion of students (especially in the experimental sample) are ranked in the lowest category. For example, in 2015 only 23% of all students (and 14% of randomized expansion sample students) met the learning expectations for their grade in math.

Teachers' pedagogical practices were observed in the subsample of 340 schools at the end of the 2017 school year. In addition, many teachers who had left these 340 schools to go to other schools in 2017 were followed and observed in their new schools. The observers assessed eight pedagogical skills of these teachers (see Table 1). These measures of pedagogical skills, and the rubric used, were designed by experts at Peru's Ministry of Education.

Table 1: Description of the Pedagogical Skills on which Teachers Were Observed

Pedagogical Skill	Description
Lesson Planning	The session's purpose is stated explicitly, in a way that students can understand. Activities are planned and aligned with the stated purpose. The session is closed referring to its purpose.
Time Management	Almost all time is allocated to pedagogical activities. Routines, transitions, and interruptions are well managed. Students know the routines and require little teacher assistance to do them.
Promotion of Students' Critical Thinking	The activities promote analysis and reasoning. Most of the questions are open ended and students are given time to delve into them.
Promotion of Students' Participation	The teacher succeeds in getting students involved and actively participating, incorporating their opinions, ideas, and interests into the session. Students can influence the class dynamics.
Provision of Oral Feedback	The teacher pays attention to the difficulties, doubts, and errors of the students, encouraging them to develop their own answers (through questions or hints), helping them to improve their understanding of the subject and advancing in their learning process. The teacher gathers evidence of the students' progress.
Provision of Written Feedback	The teacher assesses the students' work, helping them to see how to achieve what is expected of them.
Quality of Relations between Teacher and Students	Relationships in the classroom are respectful. The class sessions possess a warm environment.
Management of Students' Behavior	The teacher employs positive strategies to promote and reinforce good behavior of students, who autoregulate. An environment that promotes learning is facilitated. Bad behavior is very rare.

Several studies have found that these pedagogical skills predict student's academic success.[9] We also construct an overall index by standardizing and then averaging these eight skills.

[9] For example, Akpur (2020) finds a link between student learning and promotion of critical thinking, Stronge, Ward and Grant (2011) find a similar link for effective use of class time (which is related to lesson planning), Gage et. al. (2018) and Wisniewski, Zierer and Hattie (2020) identify provision of feedback as a mediator on student learning, Allen et. al (2013) and Fauth et. al. (2019) provide evidence in favor of promoting a positive

3. Framework and Treatment Effects

This section presents the empirical framework used in this paper. It starts by defining treatment effects for teacher skills and student learning for the context of the APM program. It then explains which treatment effects can be estimated with the available data, and then provides lower bounds for those that cannot be estimated. For details, see Appendix B in Castro et al. (2024a).

3.1 Four Types of Teachers

The Angrist, Imbens and Rubin (1996) framework divides the population of interest into *always takers*, who can always obtain the treatment, *never takers*, who can always avoid the treatment, and *compliers*, who follow their assigned treatment. Strictly speaking, these classifications are based on behavior, and do not imply any assumptions about preferences.

In the APM context, changes in treatment status occur via teacher turnover (teachers switching schools). Part of this turnover may be driven by the presence of the program, but part may also occur for reasons other than APM. If turnover is in part due to the program, it is reasonable to assume that such teachers have preferences regarding APM. We propose a framework that allows differences in preferences for APM to explain at least some teacher turnover, but we do not want turnover to be explained only by these preferences; teachers may switch schools for reasons completely unrelated to APM.

This requires changing the "traditional" classification of the population. For example, the traditional Angrist, Imbens and Rubin (1996) framework classifies a teacher moving from an APM school to a non-APM school as a *never taker*. If we assume that this is driven by a strong preference against APM, and ascribe that preference to *never takers*, we exclude the possibility that this move would have occurred even in the absence of APM.

To allow for teacher turnover that is unrelated to APM, we divide the population of teachers in the multigrade monolingual schools into four groups. First, we divide teachers into those who are relatively indifferent to APM and those with strong preferences for or against it. We further separate the latter group into *likers* (L) and *dislikers* (D). Likers are those teachers who like the program enough so that they would make a strong enough effort to secure a position in an APM school if they are not already working in one. According to the application process described in subsection 2.1, this can be done, for example, by giving a high ranking to schools that are remote or otherwise unappealing to most teachers, but that

emotional climate. Stronge, Ward and Grant (2011) and Fauth et. al. (2019) highlight the benefits of monitoring and managing student behavior. Table A9 shows that our index is positively correlated with student learning.

have the APM program. Conversely, dislikers are those teachers who dislike the program enough so that they would make a strong enough effort to secure a position in a non-APM school if they are not already working in one. Finally, we divide teachers indifferent to APM into those who have a strong enough preference for moving, but for reasons unrelated to the program, that they exert sufficient effort to move to a new school, whom we call *movers* (M), and those who remain in their schools, whom we call *remainers* (R).[10] We allow the impact of APM to differ by teachers, so we define δ^L, δ^D, δ^M and δ^R as the average effects on teacher skills of one year of APM on likers, dislikers, movers, and remainers, respectively.

Since all 6,218 randomized expansion schools followed their random assignment in 2016, all teachers in those schools had no choice regarding participation in APM in year 1.[11] We assume that teachers' behavior in the following years is characterized as follows: (i) by definition, all likers assigned to non-APM schools in year 1 move to an APM school in year 2, and all dislikers assigned to APM schools in year 1 move to a non-APM school in year 2; (ii) a fixed proportion of likers switch from one APM school to another APM school every year; (iii) a fixed proportion of dislikers switch from a non-APM to another non-APM school every year; (iv) the number of teacher positions in APM and non-APM schools is fixed; and (v) our 6,218 multigrade monolingual schools are a representative sample of a larger system of multigrade monolingual schools within which most of the teachers remain, and teacher transitions in and out this system do not affect the proportions of likers, dislikers, movers and remainers in this system.[12] As explained below, these proportions are a function of the scale of the program since teachers compete for teaching positions, and the competition to move to, or move out of, an APM school depends on the proportion of schools that are APM schools.

Comparing our four groups of teachers with the "traditional" classification above, *likers* and *dislikers* would be classified as *always takers* and *never takers,* respectively, and *remainers* can be classified as *compliers*.[13] The key difference is the addition of *movers*, whose behavior is consistent with that of any of these three traditional groups. If a mover does not change treatment status after changing schools, he or she could be seen as a *complier* according to the traditional classification. Yet if this teacher had moved from an APM school to a non-APM school he or she would be classified as a *never taker*, and if he or

[10] As almost all other studies do, we assume that there are no "defiers". Such teachers would move to a non-APM school in year 2 if they were assigned to an APM school in year 1, or move to an APM school in year 2 if assigned to a non-APM school in year 1, because they want to defy their random assignment.
[11] When teachers learned of their random assignment for 2016 it was too late to switch schools in that year.
[12] Administrative data show that, in any given year, about 10% of teachers move out of multigrade monolingual schools into other schools, and another 10% leave the public education system.
[13] All followed their assignment in year 1, so likers (dislikers) are always (never) takers only in years 2 and 3.

she had moved from a non-APM school to an APM school, he or she would be considered an *always taker*. Moreover, since movers do not take APM into account when changing schools, they always have a probability between 0 and 1 (and never equal to 0 or 1) of moving to an APM (or non-APM) school after year 1, which is not the case for any of the three "traditional" groups.

3.2 Treatment Effects for Teacher Skills.

The skill of teacher j at the end of year t, denoted by y_j^t, is assumed to be a linear function of his or her skills in the previous year (y_j^{t-1}), the skill gained from one more year of experience (λ_j), and whether he or she is treated (coached) in year t (T_j^t). The average treatment impact can vary by the four teachers types (remainers (R), likers (L), dislikers (D) and movers (M)). General depreciation of teaching skills can be included in λ_j. Equation (1) provides the general expression of y_j^t for year t, and equation (2) shows the specific expression for year 1:

$$y_j^t = y_j^{t-1} + \lambda_j + \delta^k T_j^t, \quad \text{for } k = R, L, D, M \qquad (1)$$

$$y_j^1 = y_j^0 + \lambda_j + \delta^k T_j^1 = \theta_j^1 + \delta^k T_j^1, \quad \text{for } k = R, L, D, M \qquad (2)$$

where θ_j^1 is convenient notation for $y_j^0 + \lambda_j$.[14]

In year 2, there may be interactions (denoted by $\gamma_{1,2}^k$) of the coaching in years 1 and 2:

$$y_j^2 = y_j^1 + \lambda_j + \delta^k T_j^2 + \gamma_{1,2}^k T_j^1 T_j^2, \quad \text{for } k = R, L, D, M \qquad (3)$$

$$= \theta_j^2 + \delta^k (T_j^1 + T_j^2) + \gamma_{1,2}^k T_j^1 T_j^2$$

The second line substitutes out y_j^1 using (2), and θ_j^2 denotes $\theta_j^1 + \lambda_i = y_j^0 + 2\lambda_j$. If, for example, the second year's impact of coaching is less than that of the first year, then the interaction term $\gamma_{1,2}^k$ is < 0. Also, $\gamma_{1,2}^k$ can include depreciation of teacher skills produced by the program.

For year 3, further interaction effects are needed. The equation for y_j^3 is:

[14] An implicit assumption in equation (1), and thus of equations (2) – (4), is that there are no peer effects: the skills of teacher j are not affected by whether fellow teachers have been coached. It is not possible to check this assumption with our data, yet there are three reasons why it is unlikely that a coached teacher will have sizeable impacts on the skills of other teachers in the same school. First, about 20% of the schools in the test score data, and 49% in the teacher skill data, have only one teacher; for these schools peer effects are not possible. Second, almost all schools that have more than one teacher have only two or three teachers, and they all teach different grades. For example, one teacher teaches grade 1-3 and another teaches grades 4-6. Third, coaching is generally teacher-specific, addressing the pedagogical weaknesses of a specific teacher and the needs of that teacher's students; other teachers are likely to have different pedagogical weaknesses and students with different needs; this further reduces opportunities for peer effects. If peer effects do occur, such a SUTVA violation would lead to underestimation of ITT effects, so our ITT estimates would be lower bounds for the true ITT parameters.

$$y_j^3 = y_j^2 + \lambda_j + \delta^k T_j^3 + \gamma_{1,2}^k (T_j^1 T_j^2 + T_j^1 T_j^3 + T_j^2 T_j^3) + \gamma_{1,2,3}^k T_j^1 T_j^2 T_j^3, \quad \text{for } k = R, L, D, M \qquad (4)$$
$$= \theta_j^3 + \delta^k (T_j^1 + T_j^2 + T_j^3) + \gamma_{1,2}^k (T_j^1 T_j^2 + T_j^1 T_j^3 + T_j^2 T_j^3) + \gamma_{1,2,3}^k T_j^1 T_j^2 T_j^3$$

where the second line uses (3) to substitute out y_j^2, and $\theta_j^3 = \theta_j^2 + \lambda_i = y_j^0 + 3\lambda_j$. Note that the interaction effect for any combination of two years of coaching is assumed to be the same, regardless of which two years they are; allowing for different interaction effects for each possible pair of years would do little beyond complicating the notation. The triple interaction $\gamma_{1,2,3}^k$ can include depreciation of the skills of teachers who are coached for all three years.

For the APM program, three standard treatment effects can be defined for teacher skills. The first is the average treatment effect (ATE), APM's impact on the average teacher (when all teachers are treated, i.e. receive coaching). The counterfactual is that no teachers are treated, or equivalently that the program does not exist. ATE for year t is defined as:

$$ATE_{tchr}(t) \equiv E[y_1^t - y_0^t] = E[y_1^t] - E[y^t | \text{ No program exists}] \quad (5)$$

where the "tchr" subscript indicates that the treatment effect refers to teachers' skills. For y, the superscript is still years since the program started, but subscripts indicate potential outcomes (1 = treated, 0 = not treated). Implicit in this definition is that the two potential outcomes in year t (y_1^t and y_0^t) maintain the same potential outcome status (treated or not treated) since year 1, so a teacher who is treated in year 1 is treated for all years between 1 and t, and a teacher who is not treated in year 1 is not treated for all years between 1 and t. The population of teachers for which this treatment effect is defined is all teachers who were teaching in multigrade monolingual schools in Peru in year 1.

A more specific example of equation (5) is for year 2 (t = 2), which is the only year for which teacher skill data are available. This can be expressed as:

$$ATE_{tchr}(2) \equiv E[y_1^2 - y_0^2] = 2\bar{\delta} + \bar{\gamma}_{1,2}$$

where $\bar{\delta} = \delta^R p^R + \delta^L p^L + \delta^D p^D + \delta^M p^M$, $\bar{\gamma}_{1,2} = \gamma_{1,2}^R p^R + \gamma_{1,2}^L p^L + \gamma_{1,2}^D p^D + \gamma_{1,2}^M p^M$, and p^k is the proportion of type k teachers. Appendix B in Castro et al. (2024a) gives expressions for $ATE_{tchr}(1)$ and $ATE_{tchr}(3)$.

Next consider the intention to treat (ITT) effect. This is the program's impact on skills in year t of teachers randomly assigned to APM schools in year 1, regardless of the school

they were in (APM or non-APM) in later years. The counterfactual is random assignment to a non-APM school in year 1, regardless where they taught in later years. It is defined as:

$$ITT_{tchr}(t) \equiv E[y^t|\ R_{tchr,\ year\ 1} = 1] - E[y^t|\ R_{tchr,\ year\ 1} = 0] \quad (6)$$

$R_{tchr,\ year\ 1}$ refers to the teacher's school in year 1, which can differ from his or her school in year t. An example of equation (6) is for year 2, the year with teacher skill data:[15]

$$\begin{aligned} ITT_{tchr}(2) &\equiv E[y^2|\ R_{tchr,\ year\ 1} = 1] - E[y^2|\ R_{tchr,\ year\ 1} = 0] \\ &= \bar{\delta} + p^R(\delta^R + \gamma^R_{1,2}) + p^L\gamma^L_{1,2} + p^M\tau\gamma^M_{1,2} \end{aligned}$$

where τ is the proportion of teacher positions in APM schools among the population of all monolingual multigrade schools. The intuition is that $\bar{\delta}$ is the effect of the first year, when all teachers follow their random assignment, and the other terms are the effects on the teachers treated in the second year (remainers, likers, and the movers who randomly end up in APM schools in year 2). The counterfactual for remainers is being in a non-APM schools for both years, while the counterfactual for likers, and for movers who randomly (with probability τ) end up in an APM school in year 2, is being in a non-APM school in year 1 and an APM school in year 2.

A final important point is that, unlike $ATE_{tchr}(2)$, $ITT_{tchr}(2)$ depends on τ. In a "small-scale" RCT, τ would be almost zero and so could be ignored, but in an "at-scale" RCT τ will be larger and will affect $ITT_{tchr}(2)$. The intuition is that a proportion τ of movers in APM schools in year 1 will also be in APM schools in year 2, which "turns on" the interaction effect from two years of coaching; if the proportion of APM schools had been very small, very few movers who moved into APM schools in year 2 would have been treated in year 1.

In addition, there is a more subtle impact of τ on $ITT_{tchr}(2)$: it determines the level of competition among "potential likers" to move into APM schools, and similarly the extent of competition among "potential dislikers" to move into non-APM schools. This will ultimately determine the proportions of teachers who are actual likers and dislikers, and thus the proportions of teachers who are remainers and movers. However, if there are no likers or dislikers, then the value of τ would not affect the proportions of remainers and movers.

[15] Note a slight abuse of notation: "R" is used in two different ways. If it is "normal" size (not a superscript) it indicates a school's *random* assignment, but if it is a superscript it denotes *remainer* teachers.

Another treatment effect that is often estimated for randomized control trials is a local average treatment effect (LATE).[16] It is defined only for a binary treatment variable, but the APM treatment variable can have more than two values since teachers can switch schools: the treatment can be 0, 1, 2 or 3 years. Angrist and Imbens (1995) extended LATE to non-binary treatments, which they call an average causal response (ACR). The general definition is:

$$ACR_{tchr}(t) \equiv \sum_{s=1}^{t} E[y_s^t - y_{s-1}^t | T_1^t \geq s > T_0^t] \frac{Prob[T_1^t \geq s > T_0^t]}{\sum_{r=1}^{t} Prob[T_1^t \geq r > T_0^t]} \qquad (7)$$

where T_0^t is the (potential) number of years of coaching up through year t for teachers assigned to non-APM schools in year 1, and T_1^t is the (potential) years of coaching up through year t for a teacher assigned to an APM school in year 1.[17] The subscripts on y indicate the value of y given a (potential) number of *years* of treatment (which varies from 0 to 3), not the value of y given a binary "treated or not treated" variable, in contrast to the definition of $ATE_{tchr}(t)$.

Consider equation (7) for year 2, the only year with teacher skill data:

$$ACR_{tchr}(2) \equiv E[y_1^2 - y_0^2 | T_1^2 \geq 1 > T_0^2] \frac{Prob[T_1^2 \geq 1 > T_0^2]}{Prob[T_1^2 \geq 1 > T_0^2] + Prob[T_1^2 = 2 > T_0^2]}$$

$$+ E[y_2^2 - y_1^2 | T_1^2 = 2 > T_0^2] \frac{Prob[T_1^2 = 2 > T_0^2]}{Prob[T_1^2 \geq 1 > T_0^2] + Prob[T_1^2 = 2 > T_0^2]}$$

$$= [\bar{\delta} + p^R(\delta^R + \gamma_{1,2}^R) + p^L\gamma_{1,2}^L + p^M\tau\gamma_{1,2}^M]/[1 + p^R] = ITT_{tchr}(2)/[1 + p^R]$$

The intuition behind this equation is the following. The term $E[y_1^2 - y_0^2 | T_1^2 \geq 1 > T_0^2]$ is the impact on teacher skills of receiving one year of treatment, relative to having zero years of treatment, as indicated by the subscripts on the y terms, for teachers who would have had at least one year of treatment by year 2 if assigned to an APM school in year 1 ($T_1^2 \geq 1$), but would not have been treated by year 2 if assigned to a non-APM school in year 1 ($T_0^2 < 1$). Of the four teacher types, this includes all remainers and dislikers, and movers who randomly switched to a non-APM school in year 2 (for whom $T_0^2 = 0$ and $T_1^2 = 1$). The term $E[y_2^2 - y_1^2 | T_1^2 = 2 > T_0^2]$ is the impact on teacher skills of receiving a *second* year of the treatment, *relative to having one year of treatment*, as indicated by the subscripts on the y terms, for teachers who would have had two years of treatment in year 2 if assigned to an APM school

[16] For the APM context, there is no ATT (average treatment effect on the treated) because ATT requires that some teachers assigned to the treatment ($R_{tchr,\ year\ 1} = 1$) are never treated. Such teachers do not exist in the APM context because all the teachers who were randomly assigned to the APM schools were treated in year 1.

[17] For the general case, possible values for both T_0^t and T_1^t are integers from 0 to t. Yet, for the APM program, all teachers followed their random assignment in year 1, so possible values for T_0^t are 0 to t-1, and for T_1^t are 1 to t.

in year 1 but only zero or one year of treatment in year 2 if assigned to a non-APM school in year 1. This includes all remainers, all likers, and movers who randomly switched to APM schools in year 2 (for whom $T_0^2 = 1$ and $T_1^2 = 2$). Turning to the sums of the probabilities in the denominators, $\text{Prob}[T_1^2 \geq 1 > T_0^2]$ is the probability that a teacher is a remainer, a disliker, or a mover who randomly switches to a non-APM school in year 2, and $\text{Prob}[T_1^2 = 2 > T_0^2]$ is the probability that a teacher is a remainer, a liker, or a mover who randomly switches to an APM school in year 2. Their sum is greater than 1; remainers are "counted twice" since they are included in both probabilities. Likers, dislikers, and movers are "counted" only once.

In effect, $ACR_{tchr}(2)$ is an average of: a) the (average) impact on teacher skills of going from no treatment to one year of treatment for remainers, dislikers, and those movers who randomly move to a non-APM school in year 2; and b) the (average) impact on those skills of going from one to two years of treatment for remainers, likers, and the movers who randomly move to APM schools in year 2. Thus, $ACR_{tchr}(2)$ is the average of the impact on teacher skills for each additional year of treatment due to random assignment in year 1 to an APM school, with remainers getting "double weight" since that assignment raises their years of treatment by two years, but for all others that assignment raises years of treatment by only one year. Importantly, note that, for any t, $ACR_{tchr}(t)$ is a *per year* (*not* a cumulative) impact, averaging over years of treatment induced by schools' random assignment to APM in year 1. The cumulative effect is $ACR_{tchr}(t)$ multiplied by the years of coaching induced by a school's random assignment to APM (the denominator in (7)): this equals $ITT_{tchr}(t)$. A final aspect of $ACR_{tchr}(2)$ to note is that, like $ITT_{tchr}(2)$, it is a function of τ, since its numerator is $ITT_{tchr}(2)$.

The three treatment effects discussed so far focus on particular teachers, and so they follow teachers who move to other schools. But many teacher training or coaching programs focus on particular schools, so it is useful to define treatment effects for the teachers currently in the schools that implemented APM.

There are two possibilities for treatment effects that focus on schools.[18] The first is an average treatment effect (ATE) on teacher skills for those schools, where the counterfactual is no program at all, which we denote as ATE_{sch}. This is defined as follows for year t:

$$ATE_{sch}(t) \equiv E[y^t | R = 1] - E[y^t | \text{Program does not exist}] \quad (8)$$

As above, consider again the specific case of year 2, the only year with teacher skill data:

[18] $ACR_{sch}(t)$ is not well defined since teachers who move into the 6,218 schools have no instrumental variable.

$$ATE_{sch}(2) = (2\delta^R + \gamma^R_{1,2})p^R + (\delta^L(1+\tau) + \gamma^L_{1,2}\tau)(p^L/\tau) + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M(\mu/\tau)$$
$$+ \bar{\theta}^{2,L}p^L((1-\tau)/\tau) - \bar{\theta}^{2,D}p^D + \bar{\theta}^{2,M}p^M((\mu/\tau) - 1)$$

where μ is the proportion of all movers who move to an APM school in year 2 or year 3, and the $\bar{\theta}^{2,k}$ terms are averages of θ_j^2 for year 2 for type k teachers.[19] The first line of $ATE_{sch}(2)$ is the "direct" treatment effect and the second is a "composition" effect, which accounts for differences in average θ between likers, who move into APM schools in year 2, and dislikers, who move out of APM schools in year 2 (and also accounts for changes in the distribution of movers across the two types of schools, who compete with likers to get into APM schools and with dislikers to get into non-APM schools). Note that $ATE_{sch}(2)$, and more generally $ATE_{sch}(t)$ with $t \geq 2$, also depends on τ. Intuitively, τ determines the proportions of likers and movers in APM schools (and of dislikers and movers in non-APM schools), yet this is no longer the case if there are no likers or dislikers, as explained below.

The second treatment effect for teacher skills that focuses on schools is ITT_{sch}; it is similar to ATE_{sch} except that the counterfactual is the skills of teachers in non-APM schools:

$$ITT_{sch}(t) \equiv E[y^t|\ R = 1] - E[y^t|\ R = 0] \qquad (9)$$

For year 2, this is:

$$ITT_{sch}(2) = (2\delta^R + \gamma^R_{1,2})p^R + (\delta^L(1+\tau) + \gamma^L_{1,2}\tau)(p^L/\tau) + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M(\mu/\tau)$$
$$- [\delta^D p^D(\tau/1-\tau)) + \delta^M\tau p^M((1-\mu)/(1-\tau))]$$
$$+ \bar{\theta}^{2,L}(p^L/\tau) + \bar{\theta}^{2,M}p^M(\mu/\tau) - [\bar{\theta}^{2,D}(p^D/(1-\tau)) + \bar{\theta}^{2,M}p^M((1-\mu)/(1-\tau))]$$

The first two lines are the (net) treatment effect; the last is the composition effect. As with $ATE_{sch}(t)$, $ITT_{sch}(t)$ depends on the proportion of schools that are treated (τ) when $t \geq 2$.

3.3 Treatment Effects for Student Learning

Next, consider treatment effects on student skills. Assume that the skill (measured by a test score) of student i at the end of year t, denoted by s_i^t, is determined by his or her skill at the end of the previous year (s_i^{t-1}) and the skills of his or her teacher in year t (y_j^t), where j is the teacher that student i had in year t, and π is the impact of teacher skill on student skills:

[19] To see where the μ/τ term comes from, note that the number of teaching positions in a school rarely changes. If the number of those positions is fixed in all schools, this definition of μ (where μ is determined by the application process that also determines the proportions of teachers who are likers, dislikers, movers and remainers; see subsection 2.1), implies that, among all teachers in APM and non-APM schools, the proportion who are movers in APM schools in year 2 or 3 is μp^M. Focusing on APM schools, this proportion must be divided by τ, yielding $(\mu/\tau)p^M$. Similar derivations show the proportion of movers in non-APM schools in year 2 or 3 is $[(1-\mu)/(1-\tau)]p^M$.

$$s_i^t = \sigma s_i^{t-1} + \pi y_j^t \qquad (10)$$

Each school is randomly assigned to be either an APM (R = 1) or non-APM (R = 0) school, an assignment that is fixed over time. Analysis of student skills is simplified by the fact that few students change schools (see subsection 4.2), and each school follows its random assignment.

We define three treatment effects for student skills. The first two, ATE_{stud} and ITT_{stud}, are analogous to the two treatment effects defined for their schools (ATE_{sch} and ITT_{sch}). All three treatment effects for years 2 and 3 are complex due to several possible "histories" for students' teachers in those years. For example, in year 2 a student's teacher in an APM school could be a liker who was in an APM school in years 1 and 2, or a liker who was in a non-APM school in year 1 but in an APM school in year 2. Another example is a student in an APM school in year 3; if he or she was taught by a treated teacher in year 1 (this is certain as the student was in an APM school in year 1), and by a teacher in year 2 who had APM in year 2 but not year 1, and by a teacher in year 3 who had APM in years 2 and 3 but not year 1, he or she was exposed to four years of teacher coaching, and the cumulative learning gain from this exposure is averaged over the four years. The general definition of ATE_{stud} for year t is:

$$ATE_{stud}(t) \equiv E[s^t|\ R = 1] - E[s^t|\ \text{Program does not exist}] \qquad (11)$$

Applying this definition to year 1 yields $ATE_{stud}(1) = \pi\bar{\delta}$. Applying it to year 3 (recall that test score data exist only for 2016 and 2018) yields (see Appendix B in Castro et al. (2024a) for the derivations):

$$\begin{aligned} ATE_{stud}(3) &= \sigma ATE_{stud}(2) + \pi ATE_{sch}(3) = \sigma(\sigma ATE_{stud}(1) + \pi ATE_{sch}(2)) + \pi ATE_{sch}(3) \\ &= \sigma^2\pi\bar{\delta} + \sigma\pi[(2\delta^R + \gamma_{1,2}^R)p^R + (\delta^L(1+\tau) + \gamma_{1,2}^L\tau)(p^L/\tau) + (\delta^M(1+\tau) + \gamma_{1,2}^M\tau)p^M(\mu/\tau)] \\ &+ \pi[(3\delta^R + 3\gamma_{1,2}^R + \gamma_{1,2,3}^R)p^R + (\delta^L(2+\tau) + \gamma_{1,2}^L(2\tau+1) + \gamma_{1,2,3}^L)(p^L/\tau) + (\delta^M(1+2\tau) + \gamma_{1,2}^M\tau(2+\tau) + \tau^2\gamma_{1,2,3}^M)p^M(\mu/\tau)] \\ &+ \sigma\pi[\bar{\theta}^{2,L}p^L((1-\tau)/\tau) - \bar{\theta}^{2,D}p^D + \bar{\theta}^{2,M}p^M((\mu/\tau) - 1)] + \pi[\bar{\theta}^{3,L}p^L((1-\tau)/\tau) - \bar{\theta}^{3,D}p^D + \bar{\theta}^{3,M}p^M((\mu/\tau) - 1] \end{aligned}$$

For $ATE_{stud}(3)$, the first two lines are the treatment effect, and the last line is the composition effect. Again, for t = 2 or t = 3, $ATE_{stud}(t)$ depends on τ.

Turn next to ITT. The general definition for year t is:

$$ITT_{stud}(t) \equiv E[s^t|\ R = 1] - E[s^t|\ R = 0] \qquad (12)$$

For year 1, $ITT_{stud}(1) = ATE_{stud}(1) = \pi\bar{\delta}$, as all teachers follow their schools' random assignment in year 1. For year 3, applying the general definition yields (Appendix B in Castro et al. (2024a) gives details):

$$ITT_{stud}(3) = \sigma ITT_{stud}(2) + \pi ITT_{sch}(3) = \sigma(\sigma ITT_{stud}(1) + \pi ITT_{sch}(2)) + \pi ITT_{sch}(3)$$
$$= \sigma^2\pi\bar{\delta} + \sigma\pi[(2\delta^R + \gamma^R_{1,2})p^R + (\delta^L(1+\tau) + \gamma^L_{1,2}\tau)(p^L/\tau) + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M(\mu/\tau) - [\delta^D p^D(\tau/1\text{-}\tau)) + \delta^M\tau p^M((1\text{-}\mu)/(1\text{-}\tau))]]$$
$$+ \pi[(3\delta^R + 3\gamma^R_{1,2} + \gamma^R_{1,2,3})p^R + (\delta^L(2+\tau) + \gamma^L_{1,2}(2\tau+1) + \tau\gamma^L_{1,2,3})(p^L/\tau) + (\delta^M(1+2\tau) + \gamma^M_{1,2}\tau(2+\tau) + \tau^2\gamma^M_{1,2,3})p^M(\mu/\tau)]$$
$$- \pi[\delta^D p^D(\tau/1\text{-}\tau)) + (\delta^M 2\tau + \tau^2\gamma^M_{1,2})p^M((1\text{-}\mu)/(1\text{-}\tau))]$$
$$+ \sigma\pi[\bar{\theta}^{2,L}(p^L/\tau) + \bar{\theta}^{2,M}p^M(\mu/\tau) - [\bar{\theta}^{2,D}(p^D/(1\text{-}\tau)) + \bar{\theta}^{2,M}p^M((1\text{-}\mu)/(1\text{-}\tau))]]$$
$$+ \pi[\bar{\theta}^{3,L}(p^L/\tau) + \bar{\theta}^{3,M}p^M(\mu/\tau) - [\bar{\theta}^{3,D}(p^D/(1\text{-}\tau)) + \bar{\theta}^{3,M}p^M((1\text{-}\mu)/(1\text{-}\tau))]]$$

The first three lines are the (net) treatment effect, and the last two are the composition effect. Note again that, for t = 2 or 3, that $ITT_{stud}(t)$ depends on τ.

The third treatment effect for students is the (average) impact of an additional year of teacher coaching on student learning, averaged over all additional years of that coaching that a student experiences. In effect, this is a transfer of the ACR_{tchr} treatment effects on teacher skill onto student learning, which is complicated by the many different "histories" a student can have of treated teachers in years 2 and 3. We call these treatment effects ACR_{stud}, though they differ from ACR_{tchr} (and so differ from the Angrist and Imbens ACR effects) since students are not *directly* treated but instead are *indirectly* treated by exposure to treated teachers.

The general definition of $ACR_{students}$ in year t (1, 2 or 3) is:

$$ACR_{stud}(t) \equiv \frac{E[s^t|R=1] - E[s^t|R=0]}{E[h_{tchr}(t)|R=1] - E[h_{tchr}(t)|R=0]} \qquad (13)$$

where $h_{tchr}(t)$ is the cumulative "history" from year 1 to year t of a student's exposure to teachers with APM coaching. For example, a student in a treated school in year 2 had a coached teacher in year 1, but in year 2 the teacher could have one or two years of coaching (e.g. one for a teacher in a non-APM school in year 1), so the student's $h_{tchr}(2)$ could be 2 or 3. The expected value of $h_{tchr}(t)$ averages over the types of teachers in the school from year 1 to year t.

For year 1, $ACR_{stud}(1) = ATT_{stud}(t) = ITT_{stud}(t)$ since all teachers follow their random assignment in year 1, so $ACR_{stud}(1) = \pi\bar{\delta}$. For year 3, applying the definition in (13) yields:

$$ACR_{stud}(3) = \frac{E[s^3|R=1] - E[s^3|R=0]}{E[h_{tchr}(3)|R=1] - E[h_{tchr}(3)|R=0]}$$

$$= \frac{\text{ITT}_{\text{stud}}(3)}{[1+5p^R + (3+2\tau)p^L/\tau + (2+3\tau)p^M(\mu/\tau)] - [2\tau p^D/(1-\tau) + 3\tau p^M((1-\mu)/(1-\tau))]}$$

To understand this derivation, note that the numerator is $ITT_{stud}(3)$. The first expression in brackets in the denominator, $1 + (5p^R + (3+2\tau)p^L/\tau + (2+3\tau)p^M(\mu/\tau)$, is $E[h_{tchr}(3)| R = 1]$, the average cumulative exposure to years of teacher coaching of a student in an APM school in year 3. The "1 +" term is exposure to a coached teacher in year 1. In years 2 and 3, the probability of getting a remainer teacher is p^R, and the probabilities of getting a liker or mover teacher are p^L/τ and $p^M(\mu/\tau)$, respectively. If a student gets a remainer teacher in year 2, he or she is exposed to two more years of accumulated coaching since that teacher has had two years of coaching by year 2, and if the student gets a remainer teacher in year three he or she will get three more years of accumulated coaching, for a total of five additional years (beyond year 1). If the student gets a liker teacher in year 2, the average liker teacher will have had $(1+\tau)$ years of coaching (one in year 2 and one more for a proportion τ of those teachers in year 1), and if the student gets a liker teacher in year 3, that will expose him or her to an additional $2+\tau$ years of accumulated coaching, so overall exposure to liker teachers will provide $3 + 2\tau$ years of accumulated coaching. Finally, exposure to a mover teacher in year 2 leads to $1+\tau$ additional years of accumulated coaching, and exposure to a mover in year 3 adds another $1+ 2\tau$ (since movers move randomly every year). Similar calculations for students who randomly end up in non-APM schools in year 1 (and the next two years) lead to accumulated coaching from dislikers and movers of $2\tau p^D/(1-\tau) + 3\tau p^M((1-\mu)/(1-\tau)$.

As with $ACR_{tchr}(2)$, $ACR_{stud}(3)$ is a *per year* effect (averaged over the relevant years of exposure to coached teachers). To obtain a cumulative effect for exposure to fully coached teachers in all three years, which is a weighted average over the four teacher types (where the weights are probabilities of teacher types being in treated schools), multiply $ACR_{stud}(3)$ by 6.

3.4 Treatment Effects if No Likers or Dislikers.

The treatment effects for years 2 and 3 in subsections 3.2 and 3.3 are much simpler if there are no likers or dislikers, leaving only remainers and movers. The equations simplify to:[20]

$$ATE_{tchr}(2) = 2\bar{\delta} + \bar{\gamma}_{1,2}, \text{ where } \bar{\delta} = \delta^R p^R + \delta^M p^M \text{ and } \bar{\gamma}_{1,2} = \gamma^R_{1,2} p^R + \gamma^M_{1,2} p^M$$

$$ITT_{tchr}(2) = \bar{\delta} + p^R(\delta^R + \gamma^R_{1,2}) + p^M \tau \gamma^M_{1,2}$$

$$ACR_{tchr}(2) = [\bar{\delta} + p^R(\delta^R + \gamma^R_{1,2}) + p^M \tau \gamma^M_{1,2}]/[2p^R + p^M] = ITT_{tchr}(2)/[1 + p^R]$$

$$ATE_{sch}(2) = (2\delta^R + \gamma^R_{1,2})p^R + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M$$

[20] They follow from the results in subsections 3.2 and 3.3: $p^L = p^D = 0$ and $\mu = \tau$ if there are no likers or dislikers.

$$ITT_{sch}(2) = \bar{\delta} + (\delta^R + \gamma^R_{1,2})p^R + p^M\tau\gamma^M_{1,2}$$

Note that $ITT_{sch}(2) = ITT_{tchr}(2)$, but $ATE_{sch}(2) \neq ATE_{tchr}(2)$.

$$ATE_{stud}(3) = \sigma^2\pi\bar{\delta} + \sigma\pi[(2\delta^R + \gamma^R_{1,2})p^R + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M]$$
$$+ \pi[(3\delta^R + 3\gamma^R_{1,2} + \gamma^R_{1,2,3})p^R + (\delta^M(1+2\tau) + \gamma^M_{1,2}\tau(2+\tau) + \tau^2\gamma^M_{1,2,3})p^M]$$
$$ITT_{stud}(3) = \sigma^2\pi\bar{\delta} + \sigma\pi[(2\delta^R + \gamma^R_{1,2})p^R + (\delta^M + \gamma^M_{1,2}\tau)p^M]$$
$$+ \pi[(3\delta^R + 3\gamma^R_{1,2} + \gamma^R_{1,2,3})p^R + (\delta^M + \gamma^M_{1,2}\tau 2 + \tau^2\gamma^M_{1,2,3})p^M]$$
$$ACR_{stud}(3) = \pi\frac{\sigma^2\bar{\delta} + \left((3+2\sigma)\delta^R + (3+\sigma)\gamma^R_{1,2} + \gamma^R_{1,2,3}\right)p^R + (\delta^M(\sigma+1) + \gamma^M_{1,2}\tau(\sigma+2) + \gamma^M_{1,2,3}\tau^2)p^M}{1+5p^R+2p^M}$$
$$= \frac{ITT_{stud}(3)}{1+5p^R+2p^M}$$

Note that there are no composition effects for $ATE_{sch}(2)$, $ITT_{sch}(2)$, $ATE_{stud}(3)$, and $ITT_{stud}(3)$. Also, the absence of likers and dislikers ($p^L = p^D = 0$) implies that there are only remainers and movers, and that $\mu = \tau$ (movers are equally distributed over APM and non-APM schools since they do not compete with likers or dislikers to move into an APM or non-APM school).

3.5 What Do OLS and IV Regressions Estimate?

Most, but not all, of these treatment effects can be estimated by OLS or IV regression. We have two samples of teachers, one (imperfectly) follows the teachers who were in APM and non-APM schools in year 1 (Sample 1), and the other focuses on the teachers in the APM and non-APM schools in any given year (Sample 2). OLS regression of Sample 1 teachers' skills in year t on a constant term and a dummy variable for assignment to an APM school in year 1 yields an unbiased estimate of the $ITT_{tchr}(t)$ treatment effect.[21] For example, consider year 2:

$$\hat{\beta}^y_{1\,OLS,t=2} = E[y^2|\, R_{tchr,\,year\,1} = 1] - E[y^2|\, R_{tchr,\,year\,1} = 0]$$
$$= \bar{\delta} + p^R(\delta^R + \gamma^R_{1,2}) + p^L\gamma^L_{1,2} + \tau p^M\gamma^M_{1,2} = ITT_{tchr}(2)$$

The "1" subscript indicates Sample 1 teachers. Appendix B in Castro et al. (2024a) presents this derivation, as well as those for years 1 and 3. It also presents the derivations for the other OLS and IV estima-tors in this subsection, for all three years, and shows that OLS estimation applied to Sample 2 teachers estimates $ITT_{tchr}(t)$ (recall that $ITT_{tchr}(t) = ITT_{sch}(t)$ *if* there are no likers or dislikers).

[21] Almost all of the regressions in this paper have other explanatory variables, but since random assignment is by definition uncorrelated with these other variables, the first line in the $\hat{\beta}^y_{1\,OLS,t=2}$ equation still holds by the Frisch-Waugh theorem. Regressions without these explanatory variables (e.g. Table 6) yields very similar results.

Next, consider IV estimation using Sample 1 teachers. Let $T^{Tot,t}$ denote the number of years that a teacher has participated in the program up through year t. IV regression uses random assignment as an instrument for $T^{Tot,t}$ to estimate the (average) impact of a year of exposure to the program on teacher skills. This yields unbiased estimates of $ACR_{tchr}(t)$. For year 2:

$$\hat{\beta}_1^{y}{}_{IV,t=2} = \frac{E[y^2|R_{tchr,year\ 1}=1] - E[y^2|R_{tchr,year\ 1}=0]}{E[T^{Tot,2}|R_{tchr,year\ 1}=1] - E[T^{Tot,2}|R_{tchr,year\ 1}=0]}$$

$$= (\bar{\delta} + p^R(\delta^R+\gamma^R_{1,2}) + p^L\gamma^L_{1,2} + p^M\tau\gamma^M_{1,2})/(1 + p^R) = ACR_{tchr}(2)$$

One can also apply OLS to Sample 2 teachers, the teachers who, in any given year, teach in the schools that were randomly assigned in year 1 to be APM or non-APM schools. An OLS regression of Sample 2 teachers' skills in year t on a constant and a dummy for teaching in an APM school in year t yields an unbiased estimate of $ITT_{sch}(t)$. So, for year 2:

$$\hat{\beta}_2^{y}{}_{OLS,t=2} = E[y^2|\ R = 1] - E[y^2|\ R = 0]$$

$$= (2\delta^R + \gamma^R_{1,2})p^R + (\delta^L(1+\tau) + \gamma^L_{1,2}\tau)(p^L/\tau) - [\delta^D p^D(\tau/(1-\tau)) + p^M[\delta^M(\mu-\tau^2)/(\tau-\tau^2) + \mu\gamma^M_{1,2}]$$

$$+ \bar{\theta}^{2,L}(p^L/\tau) + \bar{\theta}^{2,M}p^M(\mu/\tau) - [\bar{\theta}^{2,D}(p^D/(1-\tau)) + \bar{\theta}^{2,M}p^M((1-\mu)/(1-\tau))] = ITT_{sch}(2)$$

In general, IV estimation cannot be used for Sample 2 teachers in year 2 since some of those teachers moved into both APM schools and non-APM schools that were not part of the initial random assignment, such as teachers working in monolingual multigrade schools in year 1 that had ECE scores above the threshold that determined eligibility for the randomized expansion (see subsection 2.3). These Sample 2 teachers have no instrument, so IV estimation cannot be done for Sample 2 teachers.

Next, consider OLS regression for student test scores, more specifically regressing those scores on a constant and a dummy indicating being in an APM school. OLS regression of students' test scores in year t on a constant and a dummy for being enrolled in an APM school in year t yields an unbiased estimate of $ITT_{stud}(t)$. For years 1 and 3 this implies that:

$$\hat{\beta}^{s}_{OLS,t=1} = E[s^1|\ R = 1] - E[s^1|\ R = 0] = \pi\bar{\delta} = ATE_{stud}(1) = ITT_{stud}(1) = ACR_{stud}(1)$$

$$\hat{\beta}^{s}_{OLS,t=3} = E[s^3|\ R = 1] - E[s^3|\ R = 0]$$

$$= \sigma^2\pi e\gamma^R_{1,2})p^R + (\delta^L(1+\tau) + \gamma^L_{1,2}\tau)(p^L/\tau) + (\delta^M(1+\tau) + \gamma^M_{1,2}\tau)p^M(\mu/\tau) - [\delta^D p^D(\tau/(1-\tau)) + \delta^M\tau p^M((1-\mu)/(1-\tau))]]$$

$$+ \pi[(3\delta^R + 3\gamma^R_{1,2} + \gamma^R_{1,2,3})p^R + (\delta^L(2+\tau) + \gamma^L_{1,2}(2\tau+1) + \tau\gamma^L_{1,2,3})(p^L/\tau) + (\delta^M(1+2\tau) + \gamma^M_{1,2}\tau(2+\tau) + \tau^2\gamma^M_{1,2,3})p^M(\mu/\tau)]$$

$$- \pi[\delta^D p^D(\tau/(1-\tau)) + (\delta^M 2\tau + \tau^2\gamma^M_{1,2})p^M((1-\mu)/(1-\tau))]$$

$$+ \sigma\pi[\bar{\theta}^{2,L}(p^L/\tau) + \bar{\theta}^{2,M}p^M(\mu/\tau) - [\bar{\theta}^{2,D}(p^D/(1-\tau)) + \bar{\theta}^{2,M}p^M((1-\mu)/(1-\tau))]]$$

$$+ \pi[\bar{\theta}^{3,L}(p^L/\tau) + \bar{\theta}^{3,M}p^M(\mu/\tau) - [\bar{\theta}^{3,D}(p^D/(1-\tau)) + \bar{\theta}^{3,M}p^M((1-\mu)/(1-\tau))]] = ITT_{stud}(3)$$

Last, consider IV estimation for student test scores. The treatment for year t is the "history" from years 1 to t of students' exposure to treated teachers, $h_{tchr}(t)$ (see subsection 3.3). Thus:

$$\hat{\beta}^{s}_{IV,t} \equiv \frac{E[s^t|R=1] - E[s^t|R=0]}{E[h_{tchr}(t)|R=1] - E[h_{tchr}(t)|R=0]}$$

This is an unbiased estimate of $ACR_{stud}(t)$. Applying this to year 1, it equals OLS since all teachers follow their random assignment in year 1:

$$\hat{\beta}^{s}_{IV,1} = \pi\bar{\delta} = ATE_{stud}(1) = ITT_{stud}(1) = ACR_{stud}(1)$$

For year 3, the IV estimate is:

$$\hat{\beta}^{s}_{IV,3} = \frac{E[s^3|R=1] - E[s^3|R=0]}{E[h_{tchr}(3)|R=1] - E[h_{thr}(3)|R=0]}$$

$$= ACR_{stud}(3) = \frac{ITT_{stud}(3)}{[1+5p^R + (3+2\tau)p^L/\tau + (2+3\tau)p^M(\mu/\tau) - [2\tau p^D/(1-\tau) + 3\tau p^M((1-\mu)/(1-\tau))]}$$

3.6 Bounds on Treatment Effects for Years 2 and 3.

As explained above, for all t for which data are available, $ITT_{tchr}(t)$ and $ACR_{tchr}(t)$ can be estimated using Sample 1 teachers, and $ITT_{stud}(t)$ and $ACR_{stud}(t)$ can be estimated using student test scores from the APM and non-APM schools. In addition, for year 1 $ATE_{tchr}(1)$, which equals $ATE_{sch}(1)$, and $ATE_{stud}(1)$ can be estimated since they equal the corresponding ITT estimands. Unfortunately, $ATE_{tchr}(t)$, $ATE_{sch}(t)$ and $ATE_{stud}(t)$ cannot be estimated for $t \geq 2$. Yet, under plausible assumptions it is possible to show that ITT estimands are lower bounds of these ATE treatment effects. Turn now to these results, focusing on ATEs for which we have data to estimate; derivations, and ATEs for which we have no data, are in Appendix B of Castro et al. (2024a).

For year 2, consider $ATE_{tchr}(2)$ and $ITT_{tchr}(2)$. Their difference is:

$$ATE_{tchr}(2) - ITT_{tchr}(2) = \delta^L p^L + (\delta^D + \gamma^D_{1,2})p^D + (\delta^M + \gamma^M_{1,2}(1-\tau))p^M$$

As long as the first year of the program does not have a negative effect on the skills of likers (i.e. $\delta^L \geq 0$) and a second year does not have a negative effect on the skills of dislikers ($\delta^D + \gamma^D_{1,2} \geq 0$) or movers ($\delta^M + \gamma^M_{1,2} \geq 0$), $ITT_{tchr}(2)$ will be a lower bound for $ATE_{tchr}(2)$.

It is less clear that $ITT_{sch}(2)$ is a lower bound for $ATE_{sch}(2)$, because these treatment effects follow schools over time, as opposed to following teachers over time, and the composition of teachers in APM and non-APM schools can change over time. More specifically:

$$ATE_{sch}(2) - ITT_{sch}(2)$$

$$= \delta^D p^D(\tau/1\text{-}\tau)) + \delta^M \tau p^M((1\text{-}\mu)/(1\text{-}\tau)) - \bar{\theta}^{2,L} p^L + \bar{\theta}^{2,D} p^D(\tau/(1\text{-}\tau)) + \bar{\theta}^{2,M} p^M((\tau\text{-}\mu)/(1\text{-}\tau)).$$

It is reasonable to assume that the two δ terms (δ^D and δ^M) are ≥ 0, but the sign of the combined effect of the $\bar{\theta}^2$ terms, which reflect changes in teacher composition, is ambiguous, even though it is reasonable to assume that all the $\bar{\theta}^2$ terms are > 0. Perhaps this combined effect is close to zero and, if negative, is smaller in absolute value than the (weighted) sum of the two δ terms, so that $ITT_{sch}(2)$ is a lower bound for $ATE_{sch(2)}$, but it could be that the sum of the composition terms is negative and larger in absolute value than the expression with the two δ terms. However, if there are no likers or dislikers then there is no composition effect (since $p^L = p^D = 0$ and $\mu = \tau$) and so $ITT_{sch}(2)$ is a lower bound for $ATE_{sch}(2)$. In particular, $ATE_{sch}(2) - ITT_{sch}(2) = \delta^M \tau p^M$, which is ≥ 0 as long as $\delta^M \geq 0$, which is plausible.

Finally, turn to student skills for year three to compare $ITT_{stud}(3)$ with $ATE_{stud}(3)$:

$$ATE_{stud}(3) - ITT_{stud}(3)$$

$$= \sigma\pi[\delta^D p^D(\tau/1\text{-}\tau)) + \delta^M \tau p^M((1\text{-}\mu)/(1\text{-}\tau))] + \pi[\delta^D p^D(\tau/1\text{-}\tau)) + (\delta^M 2\tau + \tau^2 \gamma^M_{1,2}) p^M((1\text{-}\mu)/(1\text{-}\tau))]$$

$$+ \sigma\pi[-\bar{\theta}^{2,L} p^L + \bar{\theta}^{2,D} p^D(\tau/(1\text{-}\tau)) + \bar{\theta}^{2,M} p^M((\tau\text{-}\mu)/(1\text{-}\tau))] + \pi[-\bar{\theta}^{3,L} p^L + \bar{\theta}^{3,D} p^D(\tau/(1\text{-}\tau)) + \bar{\theta}^{3,M} p^M((\tau\text{-}\mu)/(1\text{-}\tau))]$$

The first three δ terms are ≥ 0 (assuming δ^D and δ^M are ≥ 0), and the $\delta^M 2\tau + \tau^2 \gamma^M_{1,2}$ term is also ≥ 0 as long as two years of exposure to the program does not reduce the skills of movers (as long as $2\delta^M + \gamma^M_{1,2} \geq 0$). Yet the sign of the combined effect of the θ terms, which reflects changes in teacher composition, is ambiguous, even though it is reasonable to assume that all of the $\bar{\theta}^3$ terms are > 0. Yet note that if there are no likers or dislikers then there is no composition effect (since $p^L = p^D = 0$ and $\mu = \tau$) and so $ITT_{stud}(3)$ is a lower bound for $ATE_{stud}(3)$.

4. Fieldwork and Data

This section describes the fieldwork and the data. The sources of the original data are Ministry of Education (2019a, 2019b). For further information on all of the data used in this paper, see Castro et al. (2024b).

4.1 Baseline Balance

To verify that the randomization yielded balanced treatment and control groups we checked the baseline balance on several school characteristics. Table 2 shows the descriptive statistics and pairwise t-tests on the difference between control and treatment groups for those school characteristics. Our preferred specification has school district fixed effects, so balance regressions include school district fixed effects, but no other controls. (School districts are subdivisions of regions, so region fixed effects are redundant; in any case baseline characteristics are similarly balanced using region fixed effects instead of school district fixed effects.)

Table 2 shows that most covariates are balanced in the test score evaluation sample, the exceptions being the number of students and, consequently, the number of teachers (since teacher assignment depends on the number of students). While this is what one would roughly expect by chance (the joint F-test is insignificant, with a p-value of 0.152), and the teacher-student ratio (which may affect treatment outcomes directly) is balanced, we control for the numbers of students and teachers in our regressions to ensure that we do not wrongly attribute to the program any differences due to this imbalance. We also show that the results are generally robust to excluding these controls (see Table A3).

4.2 Attrition

As explained in Section 2, we use information on two sets of outcome variables: teacher-level outcomes (teachers' pedagogical skills measured in a subset of the schools) and student-level outcomes (students' test scores gathered from the country's nationwide ECE assessments).

For the teacher-level outcomes, the planned pedagogical sample consisted of 364 schools from the 6,218 the randomized expansion schools: 182 were randomly selected from the 3,797 schools randomly assigned to APM and 182 were randomly selected from the 2,421 schools randomly not assigned to APM. These 364 schools were selected to observe, in the third quarter of 2017, the pedagogical practices of the teachers who: (i) had worked in one of the 364 evaluation sample schools in 2016 (Sample 1); and (ii) had worked in an evaluation sample school in 2017 (Sample 2). The former required visiting schools not in the 364 subsample in year 2 because many Sample 1 teachers changed schools between 2016 and 2017.

Table 2. Balance Table for Experimental Sample with Test Scores and Pedagogical Skills Measures

	Experimental sample with Test Scores in 2016					Experimental sample with Pedagogical Measures				
	N	Control Mean/(SD)	N	Treated Mean/(SD)	Pairwise t-test Conditional Diff.	N	Control Mean/(SD)	N	Treated Mean/(SD)	Pairwise t-test Conditional Diff.
Variable		(1)		(2)	(3)		(4)		(5)	(6)
Math Score (in 2015)	797	513.081	1120	513.439	0.566	81	494.397	69	501.125	2.630
		(93.721)		(91.070)	[0.893]		(94.245)		(90.382)	[0.874]
Language Score (in 2015)	797	517.649	1120	520.768	-0.113	81	508.422	69	514.273	1.816
		(67.236)		(64.580)	[0.970]		(65.470)		(67.811)	[0.864]
Number of Students	1,054	48.365	1509	44.283	5.420	174	30.920	166	27.259	2.334
		(28.040)		(22.882)	[0.000]		(25.637)		(21.687)	[0.384]
Number of Teachers	1,054	2.644	1509	2.514	0.185	174	1.983	166	1.801	0.156
		(1.268)		(1.192)	[0.000]		(1.140)		(1.151)	[0.245]
Number of Sections	1,054	5.824	1509	5.843	0.006	174	5.494	166	5.175	0.226
		(0.878)		(0.763)	[0.875]		(0.942)		(1.427)	[0.111]
Teacher-Student Ratio	1,036	0.063	1490	0.064	-0.001	174	0.088	161	0.098	-0.005
		(0.030)		(0.031)	[0.536]		(0.053)		(0.083)	[0.498]
Rurality scale	1,052	2.388	1509	2.199	-0.011	173	2.491	166	2.361	0.073
		(0.774)		(0.853)	[0.687]		(0.687)		(0.756)	[0.297]
Speak indigenous language (%)	1,054	4.350	1509	4.221	-0.139	174	2.244	166	2.487	-0.280
		(18.716)		(18.972)	[0.831]		(13.734)		(15.402)	[0.839]
Poverty Rates	1,030	64.653	1497	56.110	0.537	174	57.362	166	58.173	0.653
		(19.512)		(23.456)	[0.246]		(21.635)		(20.872)	[0.605]
Ceiling Material	1,019	5.717	1468	5.811	-0.063	173	5.543	160	5.631	-0.164
		(1.372)		(1.543)	[0.263]		(1.222)		(1.353)	[0.209]
Wall Material	1,019	6.068	1468	6.113	-0.088	173	5.983	160	6.088	-0.008
		(1.311)		(1.386)	[0.114]		(1.383)		(1.334)	[0.956]
Floor Material	1,019	2.856	1468	2.854	0.017	173	2.902	160	2.881	0.116
		(0.719)		(0.705)	[0.583]		(0.826)		(0.648)	[0.185]
% Teachers with degree	1,021	0.962	1480	0.962	0.004	169	0.984	159	0.979	0.005
		(0.122)		(0.124)	[0.511]		(0.079)		(0.094)	[0.638]
Internet Access	923	0.096	1327	0.102	0.016	158	0.070	150	0.093	-0.027
		(0.295)		(0.303)	[0.218]		(0.255)		(0.292)	[0.446]
Receives Textbooks	1,040	0.717	1489	0.758	0.006	174	0.661	161	0.720	-0.064
		(0.451)		(0.429)	[0.718]		(0.475)		(0.450)	[0.201]
Receives Notebooks	1,039	0.677	1489	0.701	0.016	174	0.598	161	0.640	-0.031
		(0.468)		(0.458)	[0.378]		(0.492)		(0.482)	[0.561]
School Day Length	1,039	8.208	1490	8.104	0.008	174	8.115	161	8.217	0.016
		(1.043)		(0.737)	[0.812]		(0.930)		(0.871)	[0.870]
Electricity	967	0.520	1401	0.560	0.016	155	0.503	153	0.549	-0.044
		(0.500)		(0.497)	[0.436]		(0.502)		(0.499)	[0.409]
Water	967	0.543	1401	0.519	0.015	155	0.484	153	0.497	0.019
		(0.498)		(0.500)	[0.457]		(0.501)		(0.502)	[0.729]
Sanitation	967	0.134	1401	0.148	-0.006	155	0.142	153	0.124	0.031
		(0.341)		(0.356)	[0.664]		(0.350)		(0.331)	[0.444]
Computers per Student	971	0.442	1405	0.460	0.071	155	0.194	153	0.307	-0.253
		(3.135)		(2.666)	[0.597]		(0.646)		(1.199)	[0.026]
F-test of joint significance (F-stat)					1.318					0.965
F-test, p-value					0.152					0.520
F-test, number of observations					1,538					120

Notes: This table presents the balance between treatment (APM) and control (non-APM) schools for the full experimental sample of schools with test scores, and for the subsample with measures of pedagogical practices. Columns 3 and 6 show the coefficient of regressing the treatment on each variable, including school district fixed effects. The significance of differences in columns (3) and (6) are indicated by p-values, which are in brackets. 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 3: Attrition of Sample 1 and Sample 2 Teachers and Evaluation Sample Schools in Year 2 (2017)

	Sample 1 teachers			Sample 2 teachers			Evaluation sample schools		
	APM	Non-APM	Total	APM	Non-APM	Total	APM	Non-APM	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Original (2016)	321	341	662	355	384	739	182	182	364
Observed (2017)	219	236	455	299	341	640	166	174	340
Attrition rate (%)	0.318	0.308	0.313	0.158	0.112	0.134	0.088	0.044	0.066
Difference in attrition rates		0.010 (0.036) [0.785]			0.046 (0.025) [0.068]			0.044 (0.026) [0.091]	

APM is the treated group and Non-APM is the control group. Standard errors and p-values for tests of differences in attrition rates are shown in parentheses and brackets, respectively.

It was not possible to observe the pedagogical skills of all Sample 1 teachers (see columns (1) – (3) in Table 3). In fact, Sample 1 attrition is high. This is mainly due to out-dated information on teachers' locations when the fieldwork was planned (in March of 2017, the start of Peru's school year). The teacher location information at that time indicated that, to observe all Sample 1 teachers who were still teaching, 406 schools needed to be visited, including 104 that were not in the 364 pedagogical sample schools. During fieldwork, 91.6% (372) of these 406 schools were visited (34 in very remote areas could not be visited), but outdated information often led to situations where the teachers had moved to other schools, and by the time this was discovered it was logistically impossible to go to the schools where those teachers were working. As seen in Table 3, only 68.7% (455 out of 662) of the original Sample 1 teachers were observed in 2017. Of the 207 unobserved Sample 1 teachers, 50 (7.6% of the 662) had stopped teaching in public schools, 28 (4.2%) were in one of the 34 schools that were not visited, and 129 (19.5%) had moved to a public school that was not in the planned sample of 406 schools. Turning to Sample 2 teachers (those in the 364 evaluation sample schools in year 2), 86.6% (640 out of 739) were observed in year 2 (see columns (4) - (6) of Table 2). In this sample, attrition is due mainly to 24 pedagogical sample schools in hard-to-reach areas that could not be visited in year 2 (see columns (7) – (9) of Table 3).

Non-random attrition can lead to biased estimates, especially for Sample 1 teachers, given their high rate of attrition. Yet if the average characteristics of the missing teachers are similar for APM and non-APM teachers, which for Sample 1 would be the case if the data (on where teachers who moved were working) were outdated primarily due to random factors, then this attrition will not yield biased estimates.

To check for possible bias, we do three things. First, we compare the attrition rates of the APM and non-APM groups. Table 3 shows that the differences in attrition rates are 1.7

(Sample 1) and 4.6 (Sample 2) percentage points. Neither difference is statistically significant at the 5% level, although the Sample 2 difference is significant at the 10% level and the 4.6 percentage point difference may be a concern since it is a 41% higher (15.8% vs. 11.2%) rate.

Second, we compare several observable characteristics of (non-attrited) APM and non-APM schools and teachers. Random assignment to the program in 2016 should ensure that, before any attrition occurred, the teacher characteristics were balanced for the teachers working in the APM and non-APM schools in that year. Random assignment should also ensure that the baseline characteristics of the 364 schools in the teacher skills evaluation sample are balanced. If attrition among Sample 1 teachers is random, the characteristics of the 455 teachers in Table 3 who were observed in 2017 should be similar between those who were working in APM schools and those who were working non-APM schools in 2016.

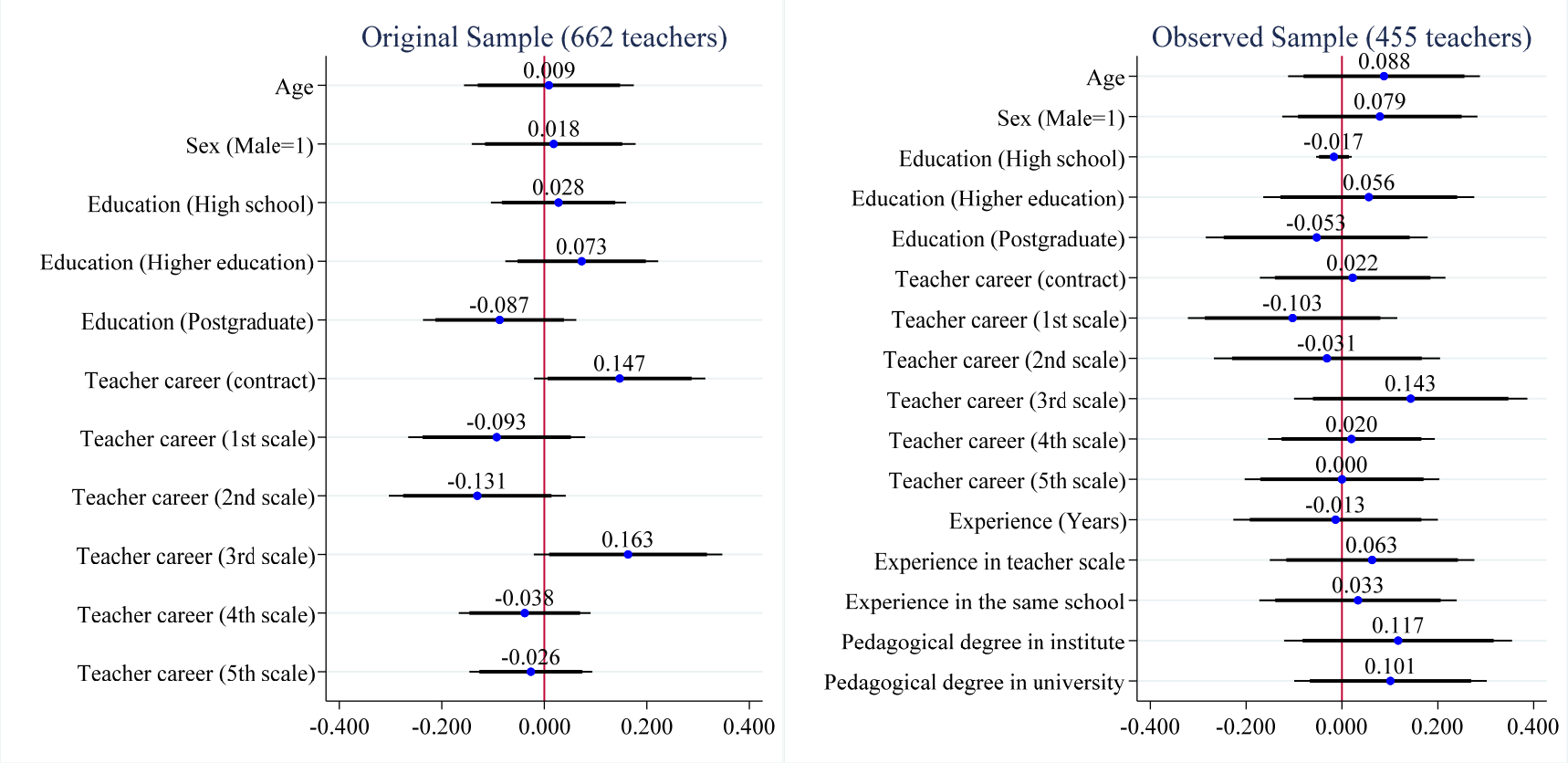
Figures 2 and 3 show that the treatment and control groups are similar in terms of the observed characteristics of: (i) the original 662 Sample 1 teachers in year 1 (2016); (ii) the subsample of 455 Sample 1 teachers who remained in that sample in year 2 (2017); (iii) the original 364 evaluation sample schools in year 1 (2016); and (iv) the subsample of 340 schools visited in year 2 (2017). Importantly, none of the (standardized) differences is very large, and none is statistically significant at the 5% level.[22]

We do not compare Sample 2 teachers' baseline characteristics in 2017 (year 2) between APM and non-APM schools to check for balance at baseline because random assignment of schools in 2016 (year 1) does not ensure such balance across these two groups of schools in year 2. In particular, if certain types of teachers self-select into APM or non-APM schools in year 2, Sample 2 teachers' baseline characteristics may be correlated with the treatment status of the schools where they worked in year 2.

Third, we used data from exams given to teachers in 2014 and 2015 that were used as part of the process by which contract teachers could become permanent civil service teachers and civil service teachers apply for promotion. We found that teachers who scored higher on those exams were less likely to move from other schools in Peru to either an APM school or a non-APM school in our 6,218 randomized expansion schools, and also that teachers who scored higher were less likely to move out of the 6,218 randomized expansion schools to other schools in Peru (see Table A10 in Castro et al. (2024a)). Most importantly, there is no relationship between these test scores and whether the teachers moved to an APM or a non-

[22] Appendix A in Castro et al. (2024a) presents further evidence that attrition is uncorrelated with treatment assignment. Table A4 shows that teachers' pre-treatment characteristics do not predict assignment to an APM school. Table A7 shows that assignment to an APM school does not predict being observed at the end of 2017.

Figure 2
Balance in Teacher Characteristics for the Original and Observed in Year 2 Teachers Who Worked in an Evaluation Sample School in 2016 (Sample 1)

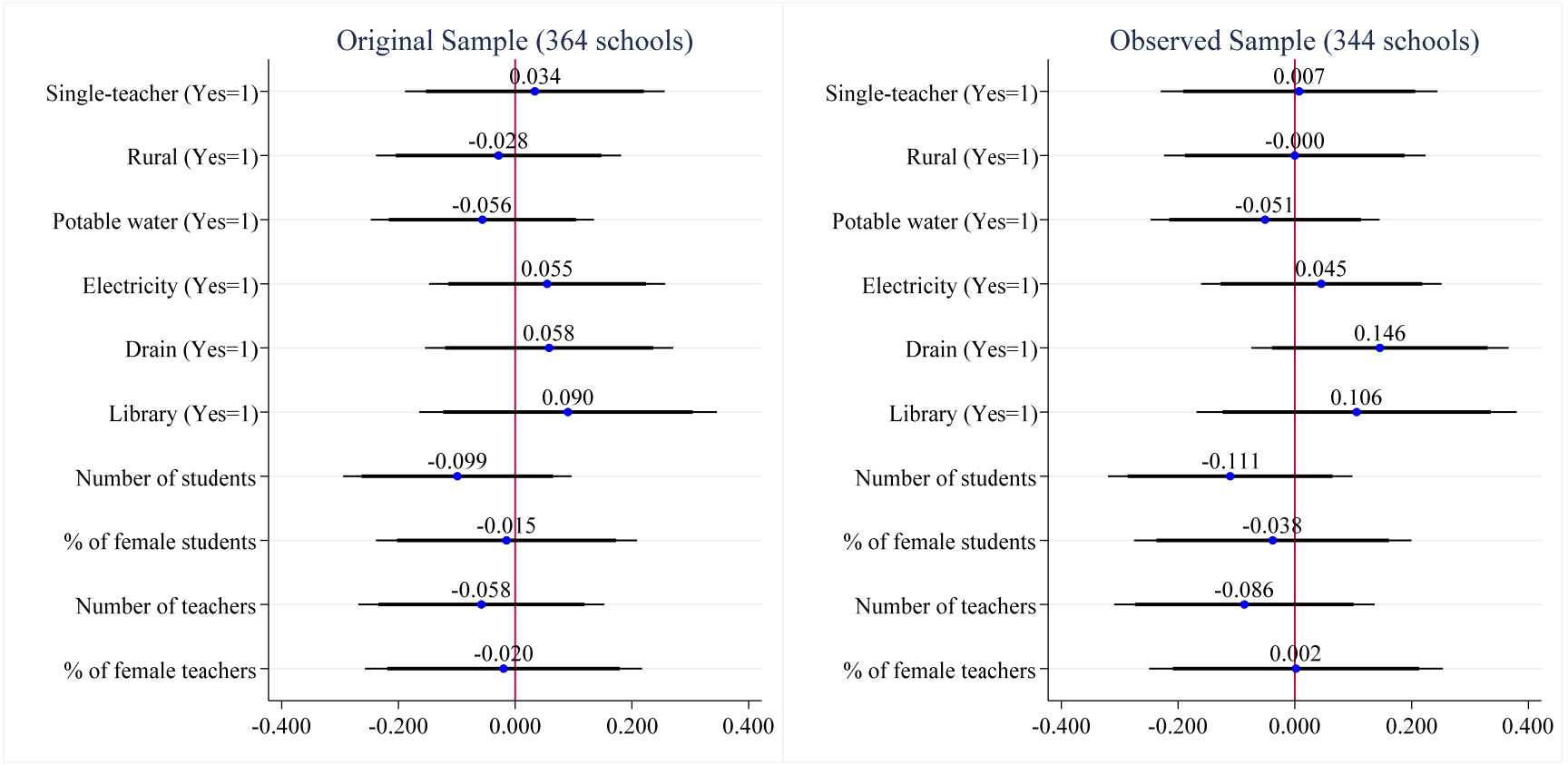


All regressions include UGEL fixed effects. Standard errors clustered at the school level.

Estimates indicate differences in the standardized characteristics of control and treatment groups. Thick and thin lines indicate 90% and 95% confidence intervals, respectively.

We do not present the differences in teacher experience and pedagogical degree for the original sample because we do not have information on those variables for the teachers that were not observed at the end of year 2.

Figure 3
Balance in School Characteristics in the Original and Observed Evaluation Sample Schools



All regressions include UGEL fixed effects.

Estimates indicate differences in the standardized characteristics of control and treatment groups. Thick and thin lines indicate 90% and 95% confidence intervals, respectively.

APM school, which shows that there is no systematic movement of better (or worse) teachers into APM or non-APM schools.

For the student-level data, there is little attrition. Using administrative data on enrollment, we found almost all the students who started in our sample in 2016 (year 1). Student turnover, unlike teacher turnover, is relatively rare in rural primary schools, especially among those targeted by APM since almost 95% are in rural areas where there are very few schools to choose from. Excluding students in their final year of primary school, and averaging over the years 2013 to 2016, only 6.9% of the students in our 6,218 primary schools in a given year were not in the same school in the next year.

4.3 Teacher Turnover and the Proportions of the Four Types of Teacher

We use administrative data on the location of teachers as well as the framework established in Section 3 to examine teacher turnover and the proportions of the four types of teacher in the sample.[23] Table 4 shows the 2016-2017 turnover behavior of Sample 1 teachers (i.e. the 12,189[24] teachers in the 6,218 randomized schools in 2016).

Table 4: Distribution of Year 1 Teachers by Their Destination School in Year 2

Treatment Arm in 2016	2016-2017 Turnover	Teachers	Percent
APM school	Stayed in the same school	4,222	63.2
	Moved to an APM school	806	12.1
	Moved to a non-APM school	1,649	24.7
	Total	6,677	100.0
Non-APM school	Stayed in the same school	2,847	62.4
	Moved to an APM school	440	9.7
	Moved to a non-APM school	1,274	27.9
	Total	4,561	100.0

By comparing the proportions of teachers in APM and non-APM schools in year 1 who moved to an APM school in year 2 (the difference between equations (A4) and (A1) in Table A5 of Castro et al. (2024a)), we estimate that $\sigma p^L = -0.024$, where σ is the proportion of likers in an APM school in a given year (e.g. year 1) who remain in the same school in the next year (e.g. year 2), rather than moving to a different APM

[23] Table A5 in Castro et al. (2024a) shows where teachers assigned to APM and non-APM schools in the randomization year end up in each type of school one year later according to their type and initial sorting.
[24] Table 4 excludes 951 teachers (7.8% of the 12,189 teachers) in the 2016 randomization sample who were not found in the administrative data in 2017; they most likely left the public education system.

school.[25] Similarly, by comparing the proportions of teachers in APM and non-APM schools who moved to a non-APM school from year 1 to year 2 (the difference between equations A5 and A2 in Table A5 of Castro et al. (2024a)), we estimate that νp^D equals -0.032, where ν is the proportion of dislikers in a non-APM school in a given year (e.g. year 1) who remain in the same school in the next year (e.g. year 2), rather than moving to a different non-APM school.

Both σp^L and νp^D are very close to 0. For σp^L to equal 0, either σ or p^L (or both) must equal 0. If $\sigma = 0$, then all likers change from one APM school to another APM school in the following year. Similarly, $\nu = 0$ implies that all dislikers already in a non-APM school in a given year move to another non-APM school the next year. Such turnover seems very unlikely since most teachers (63%) remained in the same school even before the randomized expansion of the APM program (see Table 5). By definition, likers and dislikers have strong incentives to move between schools if, in year 1, they find themselves in a school that is the opposite of their preference (likers starting in a non-APM school or dislikers starting in an APM school), but when they are placed in the school of their preferred type, we would expect turnover to be similar to what was observed in the sample before the program started, 36.6%, not 100%. Therefore, both $\sigma = 0$ and $\nu = 0$ seem very unlikely. The other option, which we consider the most realistic, is that p^L and p^D are equal to 0: there are no likers or dislikers.

The conclusion that there are no likers or dislikers is a strong claim, so we offer two additional pieces of supporting evidence. First, we analyze how teacher turnover changed over time. If there are likers and dislikers, we would expect an increased movement of teachers in the first year after the randomized expansion of APM as likers and dislikers move to the schools of their preferred type. Since schools stick to their random assignment in later years, we would expect that most of this extra turnover would occur in year 2 (2017), although some could occur in later years if some "potential" likers and dislikers are unable to move to their preferred schools in year 2. Therefore, if there are likers or dislikers, there should be a large spike in the number of teachers moving across

[25] To see how this was calculated, this definition of σ implies that the proportion of likers who move to another APM school is $1-\sigma$. Recall that μ is the proportion of movers in any school who (randomly) move to an APM school in the following year. Thus, of all teachers in an APM school in year 1, $p^L(1-\sigma) + p^M\mu$ is the proportion who move to other APM schools in year 2, and our data show that this proportion is 0.121 (see Table A5 in Castro et al. (2024a)). Similarly, the proportion of teachers in non-APM schools in year 1 who move to an APM school in year 2 is $p^L + p^M\mu$, and this proportion equals 0.097 in our data. The difference between these two proportions equals σp^L, which is -0.024 in our data. Note that this difference includes the estimates for the mentioned parameters as well as random differences in proportions that arise due to sampling. Thus small negative estimates are possible if a parameter equals 0.

treatment arms between 2016 and 2017, followed by a gradual return to regular levels of movement (from movers randomly moving between APM and non-APM schools, and likers and dislikers moving to another school of their preferred type). Table 5 shows the evolution of teacher movement across treatment arms from 2015 to 2019. There is no spike in the movement from APM to non-APM schools from 2016 to 2017; it remains at 14%, the same rate as from 2015 to 2016, and slightly less than from 2017 to 2018. A similar pattern holds for movement from non-APM to APM schools, which from 2016 to 2017 increased slightly to 12% (from 11% from 2015 to 2016) and remained at 12% from 2017 to 2018. These trends are consistent with the claim of no likers or dislikers.

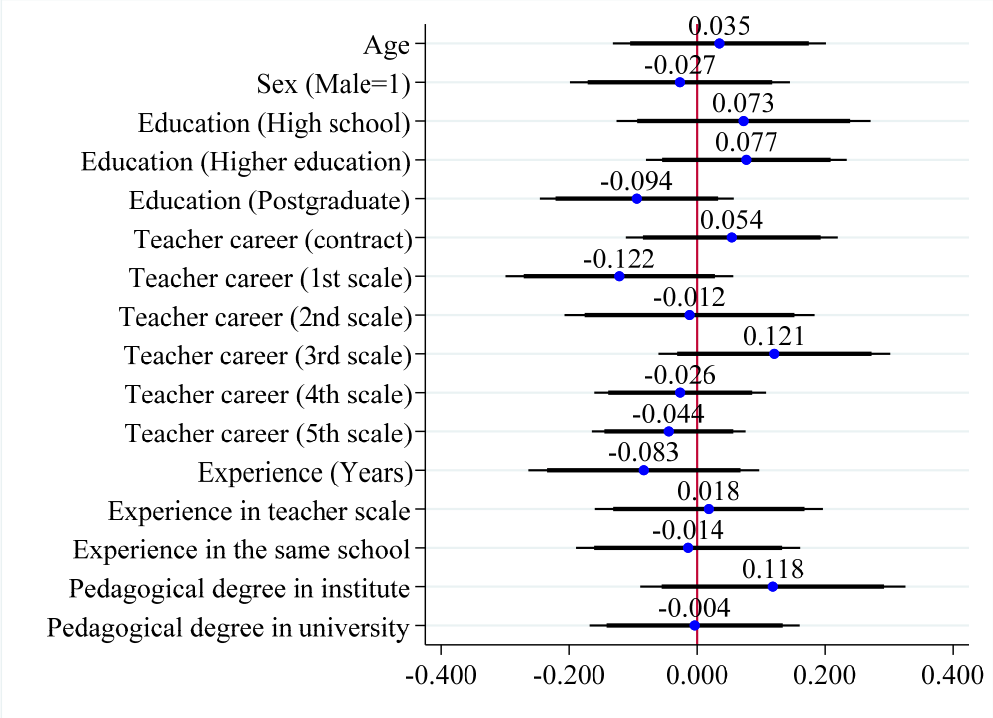
Table 5: Teacher Turnover Between APM and non-APM schools

APM Schools	2015 to 2016	2016 to 2017	2017 to 2018	2018 to 2019
Stayed in Same School	63%	65%	62%	65%
Moved to an APM School	8%	10%	10%	9%
Moved to a Non-APM School	**14%**	**14%**	**15%**	**12%**
Moved Out of Target Schools	15%	12%	14%	13%
Non-APM Schools	2015	2016	2017	2018
Stayed in Same School	63%	66%	62%	65%
Moved to an APM School	**11%**	**12%**	**12%**	**11%**
Moved to a Non-APM School	11%	10%	11%	10%
Moved Out of Target Schools	15%	12%	15%	14%
All Schools	2015	2016	2017	2018
Stayed in Same School	63%	65%	62%	65%
Moved to an APM School	10%	11%	11%	10%
Moved to a Non-APM School	12%	12%	13%	11%
Moved Out of Target Schools	15%	12%	14%	14%

Note: This table shows the year-to-year turnover status of teachers who started each two-year period in a school within one of the 6,218 randomized expansion schools.

A second piece of additional evidence for the claim of no likers or dislikers is comparisons of the characteristics of teachers who worked in the randomized pedagogical skill sample in 2017 (Sample 2). If there were likers or dislikers, one would expect the characteristics of teachers to differ between APM and non-APM schools after turnover, as likers would be only in APM schools and dislikers would be only in non-APM schools. Figure 4 shows estimates of treatment effects of APM on a wide set of teacher characteristics in the randomized expansion sample in 2017. We find no effect for any of the characteristics, suggesting that there was no systematic selection of teachers into either APM or non-APM schools, further supporting the claim of no likers or dislikers.

Figure 4: Treatment Effects on the Composition of Teacher Characteristics among the Teachers in Randomized Pedagogical Skill Sample Schools in 2017 (Sample 2)



All regressions include UGEL fixed effects.

Estimates indicate differences in the standardized characteristics of control and treatment groups. Thick and thin lines indicate 90% and 95% confidence intervals, respectively.

5. The Treatment Effects of APM

5.1 Teacher Skills

Overall teacher skills. This subsection presents estimates of $E[y^2|R_{tchr,year\,1}=1]-E[y^2|R_{tchr,year\,1}=0]$, that is estimates of $ITT_{tchr}(2)$ in equation (6), and $E[y^2|R=1]-E[y^2|R=0]$, estimates of $ITT_{sch}(2)$ in equation (9), using OLS regressions for the 455 Sample 1 teachers and the 640 Sample 2 teachers (see Table 3), respectively. We also present the estimates obtained by regressing y^2 on the predicted years of treatment, instrumented by random assignment in year 1, using Sample 1 teachers. As explained in subsection 3.5, this IV approach provides a consistent estimate of the $ACR_{tchr}(2)$ treatment effect. For all estimates, the dependent variable, y^2, is an index of pedagogical skills that averages the standardized scores of the eight indicators obtained from class-

room observations (see subsection 2.3). We present estimates with and without teacher characteristics as covariates when using Sample 1.[26] Table 6 presents these results.

Table 6: Aggregate Skill: Ordinary Least Squares (OLS) Estimates and IV Estimates

	Ordinary Least Squares Estimates			IV Estimates	
	Sample 1		Sample 2	Sample 1	
	(1)	(2)	(3)	(4)	(5)
Treatment	0.275	0.300	0.197	0.152	0.166
	(0.103)	(0.097)	(0.098)	(0.052)	(0.048)
	[0.008]	[0.002]	[0.046]	[0.003]	[0.001]
Experience	--	0.000	--	--	-0.000
		(0.009)			(0.008)
Contract teacher	--	0.139	--	--	0.133
		(0.155)			(0.138)
Teacher career Level	--	0.109	--	--	0.107
		(0.044)			(0.039)
Sex (men = 1)	--	-0.300	--	--	-0.301
		(0.095)			(0.085)
Age	--	-0.028	--	--	-0.027
		(0.009)			(0.008)
R^2	0.29	0.37	0.23	0.29	0.37
Sample Size	455	455	640	455	455

All regressions include UGEL fixed effects. Standard errors clustered at the school level are presented in parentheses, and p-values shown in brackets.

Before discussing the results, recall the claim (subsection 4.3) that our population of teachers has no likers or dislikers. Recall also (subsection 3.4) that, if there are no likers or dislikers, both $\hat{\beta}_1^y{}_{OLS,t=2}$ and $\hat{\beta}_2^y{}_{OLS,t=2}$ estimate $ITT_{tchr}(2)$, which equals $ITT_{sch}(2)$. Thus, all OLS estimates in Table 6 consistently estimate the same parameter.

The first and second columns of Table 6 present estimates of $ITT_{tchr}(2)$. The estimate in Column (1), which does not control for teacher characteristics, indicates that offering APM for two years increases teachers' pedagogical skills by 0.28 standard deviations (s.d.). The estimate in Column (2), when teacher characteristics are added as covariates, is very similar: 0.30 s.d. The estimate for $ITT_{sch}(2)$ in Column (3), 0.20 s.d.,

[26] The use of teacher characteristics as covariates is appropriate only for Sample 1 because characteristics of Sample 2 teachers can be affected by the treatment. In Table A6 of Castro et al. (2024a), we test for interactions between the treatment status and the characteristics of Sample 1 teachers. We find no evidence of heterogeneity by teacher experience, type of contract, position in the teacher career or sex. These results support the linearity assumption for the teacher skills production function in equation (1).

is somewhat lower, even though $ITT_{sch}(2)$ should equal $ITT_{tchr}(2)$. Recall that Sample 1 teachers had high rates of attrition due to difficulties finding teachers who moved; this implies that remainers are very likely overrepresented in Sample 1. In contrast, the proportions of remainers and movers in Sample 2 should correspond to their proportions in the population of teachers in the 6,218 randomized expansion schools. Thus, the Column (3) estimate is our preferred estimate of $ITT_{tchr}(2)$, which also equals $ITT_{sch}(2)$; the effect after two years on teachers' aggregate pedagogical skill of *assigning* them to an APM school in year 1 is a 0.20 s.d. increase in those skills

Our estimate that $ITT_{tchr}(2) = ITT_{sch}(2) = 0.20$ sheds some light on other parameters of interest. Recall that, in general, $ATE_{tchr}(2) \geq ITT_{tchr}(2)$, and if there are no likers and dislikers then $ATE_{sch}(2) \geq ITT_{sch}(2)$. Thus the effect of two years of APM coaching on the aggregate pedagogical practice of the average teacher, $ATE_{tchr}(2)$, and the effect of APM on the aggregate pedagogical practice of the teachers in APM schools in year 2, $ATE_{sch}(2)$, are at least as large as, and likely larger than, 0.2 s.d.

Columns (4) and (5) in Table 6 present our IV estimates of $ACR_{tchr}(2)$ using Sample 1 teachers. They show that, averaging over all years of coaching received, an additional year of coaching increases by 0.15 to 0.17 s.d. the average pedagogical skill of all teachers, but this average gives remainers a "double weight" because random assignment to an APM school induces them to obtain two years of coaching. Consistent with the fact that $ACR_{tchr}(2)$ equals $ITT_{tchr}(2)/(1+p^R)$, this IV estimate, which is a per year estimate, is somewhat larger than (half of) the Sample 1 estimate of $ITT_{tchr}(2)$, an estimate of cumulative impact over two years, in column (2).

Specific Pedagogical Skills. The discussion thus far has focused on the aggregate index of pedagogical skills, but one can also estimate $ITT_{tchr}(2)$ for each of the eight more specific pedagogical skills shown in Table 1. Table 7 shows these results. To minimize spurious statistical significance due to multiple hypothesis testing, Table 7 also presents adjusted p-values, using the Romano and Wolf (2016) stepdown method to account for multiple hypothesis testing; these are in brackets below the standard errors.

The estimates in Table 7 indicate that the biggest impact of assigning teachers to the APM program, in terms of both the size and the statistical significance of the estimated parameters, is on teachers' lesson planning; the point estimates are 0.34 s.d. for Sample 1 and 0.39 s.d. for Sample 2. There is also evidence that APM raises teachers' pedagogical skills in developing their students' critical thinking, although the statistical significance is at best only marginal after controlling for multiple hypothesis testing.

Table 7
Disaggregated Skills: Ordinary Least Squares Estimates

	(1) Lesson Planning	(2) Time Management	(3) Critical Thinking	(4) Student Participation	(5) Class Feedback	(6) Written Feedback	(7) Classroom Relationships	(8) Behavior Management
Panel A. Sample 1								
Treatment	0.338	0.083	0.248	0.162	0.199	0.136	0.069	0.114
	(0.106)	(0.108)	(0.092)	(0.103)	(0.107)	(0.096)	(0.112)	(0.098)
	[0.018]	[0.692]	[0.064]	[0.474]	[0.339]	[0.499]	[0.692]	[0.561]
N	448	450	450	450	450	448	450	450
R-squared	0.307	0.221	0.281	0.364	0.371	0.332	0.263	0.277
Panel B. Sample 2								
Treatment	0.387	-0.073	0.186	0.062	0.094	0.173	0.022	0.019
	(0.091)	(0.099)	(0.090)	(0.089)	(0.103)	(0.097)	(0.091)	(0.088)
	[0.002]	[0.917]	[0.274]	[0.917]	[0.898]	[0.428]	[0.966]	[0.96]
N	633	633	633	632	633	631	633	632
R-squared	0.245	0.171	0.200	0.260	0.277	0.236	0.209	0.238

Note: Effects are measured in standard deviations. Regressions of Panel A include the following control variables: experience, contract teacher, teacher career level, sex and age. All regressions include UGEL fixed effects. Standard errors clustered at the school level are reported in parentheses and adjusted p-values for multiple hypotheses testing are reported in brackets. We calculate the adjusted p-values using the stepdown method of Romano and Wolf (2016).

5.2 Student Learning

This subsection explores the impact of the APM coaching program on student learning, as measured by the National Student Evaluation (ECE) taken one and three years after the program began (that is, 2016 and 2018). We compare student test scores in the APM and non-APM schools in the much larger student test score sample. This sample is not restricted to the 340 schools with pedagogical practices data, but it is restricted to those schools that participated in the 2016 ECE and the 2018 ECE. As explained earlier, only schools with five or more students in the relevant grade take the ECE, so we have test scores for only 2,567 of the 6,218 randomized expansion schools.

Table 8 presents estimates of the APM coaching program's treatment effects on average ECE scores for the sample of 2,567 schools in 2016 and 2018, after one and three years of coaching. The ECE is taken at the end of the school year (which is also the end of the calendar year), so the 2016 ECE yields estimates of the APM program's impact after one year for students in grade 2. All teachers complied with their random assignment in 2016, so this is an estimate of $ATE_{stud}(1)$, the average treatment effect of one year of APM on student learning. In 2018, the ECE was conducted again, but this time it was done in grade 4, which in general contains the same students who were tested in 2016 in grade 2, except that it excludes students who repeated grade 2 or 3 (about 7-8% of students repeat each year). The 2018 ECE allows us to test for the impact of the program after three full years of implementation. Students almost always comply with treatment assignment, yet many teachers switched schools between 2016 and 2018, so we cannot estimate the average treatment effect, $ATE_{stud}(3)$ for three years. Rather, we estimate $ITT_{stud}(3)$, which is a lower bound of $ATE_{stud}(3)$ if there are no likers or dislikers.

5.2.1 Results after one year. Table 8 presents estimates of the program's treatment effects on standardized test scores for mathematics and reading comprehension.[27] Columns (1) and (5) show estimates of $ATE_{stud}(1)$ after one year of implementation, columns (2), (3), (6) and (7) show ITT and ACR estimates after three years of the program, and columns (4) and (8) present combined ACR results for years 1 and 3. While the program was designed by the

[27] Recall that ECE scores exist only for schools with five or more students in a given grade; this greatly reduces the number of schools in the student test score sample. Table A1 in Castro et al. (2024a) shows that almost all characteristics of the schools with test scores are very similar to those for the 6,218 randomized expansion schools. The baseline balance in Table 2 is for this smaller subsample of schools, which is the relevant sample for analysis.

Ministry of Education, it was implemented by each local school district (UGEL),[28] so our preferred specification, shown in this table, includes school district fixed effects, which also control for any differences in actual program implementation within each region. All Table 8 regressions also control for school size (number of teachers and students), which was slightly unbalanced at baseline. We cluster standard errors at the school level in all regressions, following Abadie et al. (2023), since the treatment is assigned at the school level.

Table 8: Results on Student Learning After One and Three Years of Coaching

	Mathematics				Reading			
	1 Year	3 Years		Combined	1 Year	3 Years		Combined
	OLS	OLS	IV	IV	OLS	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.106	0.114			0.075	0.100		
	(0.034)	(0.033)			(0.032)	(0.031)		
	[0.002]	[0.001]			[0.019]	[0.001]		
Cumulative years treated			0. 030	0.107			0.027	0.076
			(0.009)	(0.034)			(0.008)	(0.032)
			[0.001]	[0.002]			[0.001]	[0.017]
Cumulative year treated × year 3 dummy variable				-0.075				-0.049
				(0.032)				(0.029)
				[0.018]				[0.095]
Sum of above two rows				0.032				0.027
				(0.009)				(0.008)
				[0.000]				[0.001]
Coefficient on random assignment in first-stage regression			3.739				3.739	
			(0.048)				(0.048)	
F-statistic (for cumulative years treated)			6,123	5,315			6,127	5,316
F-statistic (for cumulative years treated × year 3)				3,854				3,856
Control Mean	0.003	0.004	0.004	0.004	0.004	0.003	0.003	0.004
Observations	22,198	18,261	18,261	40,459	22,199	18,275	18,275	40,474
Schools	2,547	2,053	2,053	2,547	2,547	2,053	2,053	2,547
R^2	0.142	0.182	0.184	0.143	0.162	0.168	0.169	0.153

Note: This table shows treatment effects of the coaching program on standardized student test scores. Columns 1 and 5 show the ITT effects after one year of treatment in 2016, while columns 2 and 6 show the ITT effects after three years of treatment in 2018. Columns 3 and 7 present 2SLS estimates of ACR using the random treatment assignment as an instrument for the total coaching years to which students were exposed through their teachers over the course of three years. Finally, columns 4 and 8 combined the IV regressions for years 1 and 3 (because of almost perfect compliance in year 1, IV and OLS estimates are almost identical); see the text for how to interpret the coefficients for these regressions. All specifications include school district fixed effects and control for school size (number of teachers and students), which is not balanced at baseline (See Table A3 for additional specifications). All results use standardized exam scores and can be interpreted as standard deviations. Regressions are run at the student level, with robust standard errors, clustered by school, presented in parentheses, and p-values shown in brackets.

[28] Peru's 225 school districts (UGELs) are managed by school boards, which implement education policies in their districts. Each UGEL is overseen by its Regional Education Board (Dirección de Educación Regional).

The APM coaching program has significantly positive impacts on student learning. After one year, average test scores increase by 0.106 and 0.075 standard deviations (s.d.) in math and reading comprehension, respectively. These are average treatment effects, $ATE_{stud}(1)$, and they suggest that coaching that provides regular, individualized support to teachers can be an effective policy to increase student learning. For perspective, note that the effect after one year is similar in magnitude to the median effect on learning outcomes of 234 education studies in low and middle income countries reviewed by Evans and Yuan (2022). And when compared to the median for large studies (those with over 5,000 students), the effect of the APM program after only one year is almost double that median effect (0.05 s.d.).

Table A3 in Castro et al. (2024a) shows how estimates change when using regional, rather than school district, fixed effects, and when excluding controls. Both of those changes reduce the size of the coefficient slightly, but the results are generally robust to these changes.[29] Table A3 includes another specification, column (4), that adds to the analysis the panel data available from 2010 to 2018 and adds school-level fixed effects and state-specific time trends, without any controls; its results are very close to those of main OLS specification in Table 8.

5.2.2 Results after three years. Columns (2) and (6) of Table 8 show the effects of the APM program in 2018, after three years. Recall that in 2018 the standardized test is for grade 4, so that, except for repeaters, we follow the same students observed in 2016 in grade 2 after two more years of exposure to APM. The estimated program effects, which are now ITT effects ($ITT_{stud}(3)$) and so are lower bounds for ATE ($ATE_{stud}(3)$), remain positive after three years of the program and are slightly higher (than the estimates after one year shown in columns (1) and (5)): 0.114 s.d. for math, 0.100 s.d. for reading comprehension.

These ITT results show the average effect on students learning after three years for schools that were randomly assigned to the APM program in 2016. Yet the exposure of students to treated teachers, and therefore the effective treatment dose, differs widely among APM schools as a result of teacher turnover. To estimate the impact on students of being exposed to one more year of teacher coaching, we use random assignment in 2016 to instrument students' exposure to coached teachers in each school. We have data on teachers' school assignment, so we constructed a variable that captures the intensity of coaching for the

[29] The exception is reading comprehension scores after one year of APM; they are significant only if controls are included. Yet the treatment effects after three years are robust even when excluding controls for both subjects.

teachers present in each year (since the program started) in a given school. This incorporates the coaching history of all teachers that the students had over the course of three years.[30]

We created a variable that measures the variation in the intensity of coaching received by teachers, who received either one, two or three years of coaching in the past three years; students, in turn, were exposed, over those three years, to teachers with varying years of treatment. Our constructed variable is based on the total years of coaching that each teacher received and calculates for all students the average intensity of coaching that the teachers in their school had received, for each of the three years that a student was in his or her school.

For example, students in a school A that was randomly assigned to be an APM school would have been exposed to teachers with one year of coaching in year 1 (since all schools complied with their random assignment and teachers had not yet been able to switch schools). If all teachers remain in that school the students in school A would be exposed in their second year to a teacher coached for two years, bringing their total coaching exposure to three years (one in the first year and two in the second), and similarly (if there were no teacher turnover) would be exposed to teachers with an average of three years of coaching in year 3, bringing students total exposure to six years of coaching over the three years. In contrast, students in an APM school B that experiences full teacher turnover each year would be exposed to one year of coaching in year 1, another year of coaching in year 2 (assuming that all teachers left and all new teachers had not been coached in their first year), and another year of coaching in year 3 (if all teachers once again left and all new teachers had not been coached in years 1 and 2), for a total of 3 years of coaching exposure. Students in non-APM schools that receive treated teachers can also receive exposure to some years of coaching, depending on the extent of coaching that their newly arrived teachers received previously.

We created a variable for student exposure to coached teachers up though year 3 that ranges from 0 to 6 years. The average value of this exposure variable in 2018 is 4.1 years for APM schools and 0.3 years for non-APM schools. The coefficient on this instrumented variable thus measures the effect on student test scores of an additional year of teacher coaching induced by random assignment. That is, it estimates $ACR_{stud}(3)$, the average per-year-of-coaching effect. To obtain a cumulative effect for exposure to fully coached teachers in all three years, one can multiply this coefficient by 6.

[30] Strictly speaking, we construct and average "history" over all teachers in a given school in a given year, since we cannot match students to individual teachers. Note, however, that 20% of the schools in our student test score sample had only one teacher, so for these schools we are matching students to their specific teacher.

As with the OLS estimates, our preferred specifications for both stages of the IV estimates include school district fixed effects and controls for number of teachers and number of students. Standard errors are clustered at the school level.

Columns (3) and (7) of Table 8 present the estimates of $ACR_{stud}(3)$, the per-year-of-coaching impact after three years of the APM program, on student learning, instrumenting students' exposure to teachers' accumulated coaching by schools' random assignment. They indicate that, averaging over three years of exposure, exposure to an additional year of teacher coaching raises a student's math and reading test scores by 0.030 and 0.027 standard deviations, respectively. Thus, after three years a student who had fully coached teachers (six years of exposure to coaching) would have math and reading test scores that are 0.180 and 0.162 standard deviations higher than a student not exposed to any coaching. These impacts are larger than the ITT estimates in columns (2) and (6) because they measure the impact of "full" student exposure (six years) to APM coaching over three years, while the ITT estimates compare the actual exposure to coaching of students in treated and control schools. That is, after three years the average student in an APM school was exposed to 4.1 years of teacher coaching and the average control school student was ex-posed to 0.3 years of teacher coaching, for a difference of 3.8 years. Multiplying the per year effects in columns (3) and (7) by this difference gives impacts of 0.114 for math and 0.102 for reading that, aside from rounding error, are the ITT effects measured in columns (2) and (6).

The IV estimates in Table 8 suggest sharply decreasing marginal effects to the second and third years of coaching, since the impacts on math (reading) scores are 0.106 (0.075) standard deviations for the first year of exposure to coaching, but then averaging over three years the average impact is only 0.030 (0.027) standard deviations per year of coaching. Columns (4) and (8) in Table 8 test whether these impacts are significantly different by presenting combined regression IV estimates of both the impact of the first year of coaching and the average impact over three years of coaching. This regression includes cumulated years of students' exposure to coaching (for both samples) and the same variable interacted with a dummy variable for year 3 (2018), which applies only to the year 3 data.[31] The latter variable allows the impact of the cumulated exposure to coaching to be different for year 3 observations, which means that the former estimates the impact for the year 1 observations,

[31] These two variables are instrumented by the treatment dummy variable, applied to all observations, and the same variable interacted with a dummy variable for year 3.

and it essentially reproduces the results in column (1) for math and column (5) for reading.[32] The estimates for the cumulative exposure to coaching interacted with the year 3 dummy variable are negative, which indicates that the average effect of a student's exposure to a year of teacher coaching in years 2 and 3 is smaller than the impact of the first year of such exposure; this difference is large and statistically significant for math, and somewhat smaller and only marginally statistically significant for reading (p-value = 0.095). The next row in columns (4) and (8) in Table 8 show the difference between these two estimates, which is essentially the ACR estimates for year 3 in columns (3) and (7).

There are at least four possible reasons for the declining impact of additional years of student exposure to years of teacher coaching. First, and perhaps most obvious, the increase in a teacher's pedagogical skills from a second or third year of coaching is likely to be smaller, and perhaps much smaller, than the impact in the first year, due to standard decreasing marginal returns to any input in a skills production function; such decreasing returns is allowed for in equation (4) for teacher skills in year 3. Second, students' acquired skills could depreciate over time, as indicated in equation (10) for student skills. Third, teachers' coaching skills could also depreciate over time, which is also allowed for in equation (4). Fourth, there may be decreasing marginal effects of teacher skills on student learning; however, these effects are likely less pronounced than decreasing marginal effects of coaching on teacher skills since, unlike coaching, teachers' skills are not zero at baseline (concavity in these production functions is likely more pronounced when the corresponding inputs are close to zero). Unfortunately, with the data at hand we cannot estimate the contribution of each of these factors to the declining impact of students' exposure to additional years of teacher coaching.

5.2.3 Student Skill Distribution. Teachers could respond to the treatment in various ways: for example, they could focus their efforts on the lower end of the student learning distribution to help the weaker students or, given the high stakes nature of the tests,[33] they could focus on top students by shifting resources and attention away from those who struggle. Another possibility is that they could acquire skills that help them engage with students across the entire student skill distribution. To test which part of the student grade distribution is shifting

[32] Since there was almost perfect compliance in year 1, OLS and IV results are almost identical.
[33] There are some incentive payment schemes that pay teachers bonuses according to their schools' performance on these tests. They should not affect our estimates since they apply to both APM and non-APM schools.

in response to the treatment we run a quantile regression, taking advantage of the availability of individual student test scores.

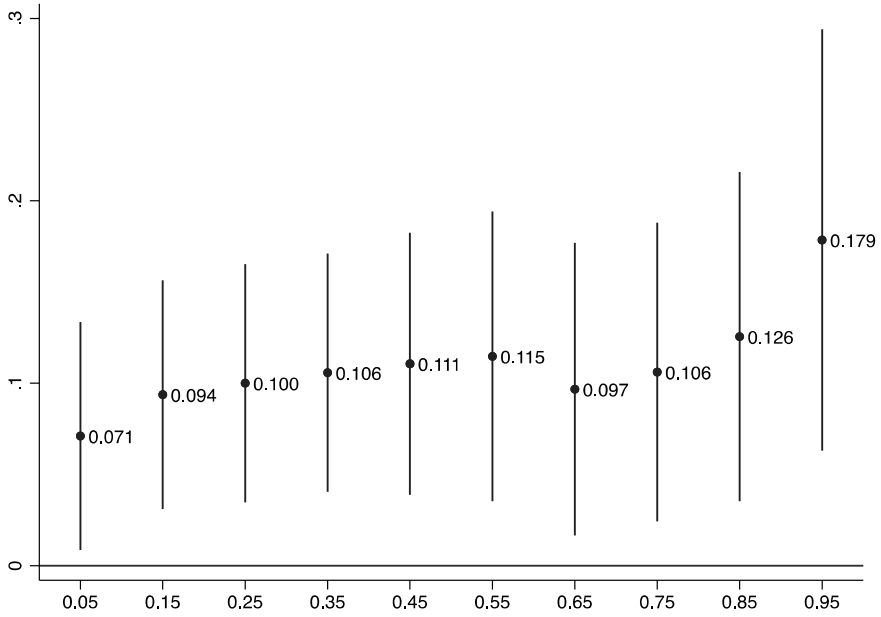
Figure 5 shows the results for quantile regressions for each decile of the student test score distribution, using our preferred specification that has school district fixed effects, after three years of the APM program (in 2018). We find that the program raised student test scores along the entire student skill distribution and we cannot reject that treatment effects are constant across all deciles. This suggests that the program, which focuses on individual teacher weaknesses, helps teachers to deal with the particular challenges their students face regardless of those students' position in the student skill distribution.

6. Concluding Remarks

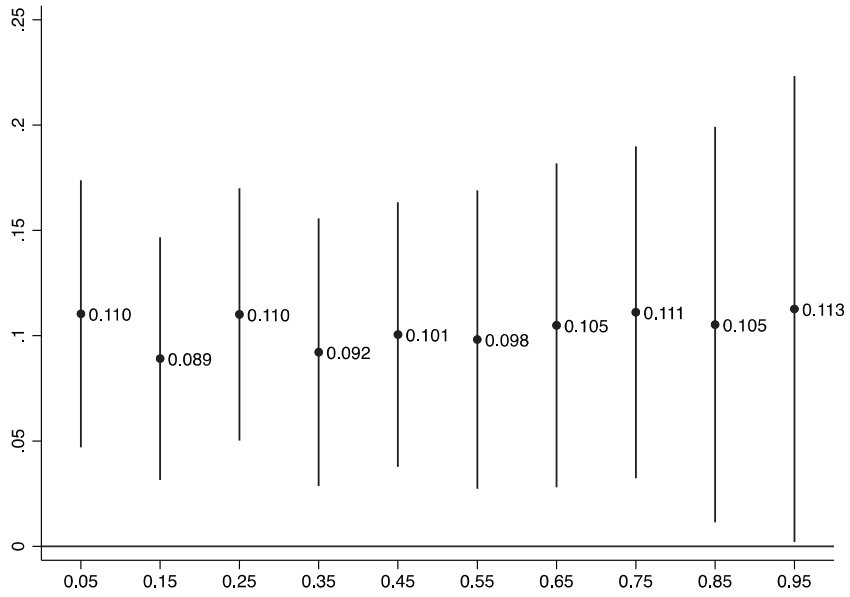
Teacher quality plays a key role in student learning, but the quality of teachers is often low, especially in developing countries. Given the vital role that human capital plays in economic growth and individuals' income and well-being, a key policy priority is to develop policies that increase teacher quality. The success of teacher training programs in raising teacher quality is, at best, mixed, but teacher coaching programs are a promising policy option.

We have estimated the effect of a large-scale teacher coaching program operating in a context of high teacher turnover in rural Peru on a broad range of pedagogical skills, and on student learning. Our study contributes to the literature on teacher training and pedagogy by addressing the issues of scale and teacher turnover as potential threats to the effectiveness of coaching, and by presenting evidence that the general pedagogical skills of the current stock of teachers can be improved. This research also contributes to the literature by developing an analytical framework that defines different types of treatment effects when teacher turnover is present and explains which treatment effects can be estimated.

Figure 5: Quantile Regression Results After Three Years of Implementation (2018)



a) Mathematics



b) Reading Comprehension

Note: T hese figures show the quantile regression coefficients for the effect of the program on standardized test scores after three years of implementation (2018) 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.

When teacher turnover is present, the success of teacher training or coaching programs can be judged from two perspectives, the impact on the teachers who were initially offered the treatment, regardless of whether they stay in their schools or move to a different school, and the impact on the teachers and students in treated schools after turnover has occurred. It is possible to estimate intent to treat (ITT) effects for the first perspective if one has a sample of teachers that follows them when they change schools, or if one has data on teacher skills from the schools that were randomly assigned to treatment and control groups *and* teacher turnover is unrelated to the program, and for the second perspective using data on teachers' skills and student learning in treated and control schools after turnover has occurred. We also show that, unfortunately, average treatment effects (ATE) cannot be estimated without bias even when turnover is unrelated to the program. However, we show that ITT estimates serve as a lower bound for ATE for the teachers who were initially offered the treatment (the first perspective). Yet from the second perspective ITTs are a lower bound for ATEs only if teacher turnover is unrelated to the program. We believe that this framework can be useful for future education evaluations carried out in contexts of high teacher turnover or, more generally, in any evaluations where treatments are offered at a cluster-level and service providers can change clusters while the intervention is still in progress.

We find that, after two years, the program has an (average) intent to treat (ITT) effect that increases teachers' pedagogical skills by 0.20 s.d.: this estimate applies to both perspectives. This effect is concentrated on two dimensions of the pedagogical practices: lesson planning and, to a lesser extent, encouraging students' critical thinking. We also estimated the ITT effect of the program on student learning and found positive effects after one year (0.075-0.106 s.d.) and after three years (0.100-0.114 s.d.) of coaching.

This research also contributes to the discussion about how to improve the pedagogical skill of teachers serving rural schools in ways that are most cost-effective. Rural schools are often located in hard-to-reach areas that tend to be avoided by teachers if they are given a choice. One potential way to improve pedagogical skills and student learning in rural schools is to offer incentives to attract more talented teachers. The rural bonus scheme in Peru pursues this objective by offering a 30% salary increase to those teachers who accept a position in a rural school. This bonus has had a small positive effect on the probability of filling a teacher vacancy but has shown no effects on learning outcomes (Castro and Esposito, 2022).

The cost of the coaching program evaluated in this study is around US$ 3,000 per teacher, per year. This is about 30% of the average annual salary of a primary school teacher

in Peru, and it is similar to the wage premium offered by the bonus program, with two important differences: coaching is only a three-year investment (not a permanent salary increase), and we have shown that it is effective for increasing student learning.

Another policy to increase pedagogical skills and student learning in rural schools is to offer incentives for (current) teachers to increase their productivity. Recent studies have shown that expensive policies based on large unconditional salary increases can reduce the number of teachers taking second jobs but have no effects on teacher productivity (de Ree et al., 2018). Pay-for-performance programs offer another alternative to improve teachers' productivity. The impact of these types of incentives has been examined in several low and middle-income countries, with mixed results. Very few studies, however, have estimated the effect of such programs in the context of a nation-wide intervention. A recent study by Bellés-Obrero and Lombardi (2022) evaluated the effect of a national pay-for-performance program implemented in 2015 in public secondary schools in Peru. The program, *Bono Escuela*, offers an additional monthly salary to the principal and teachers of the schools that rank in the top 20% of the national 8th grade student evaluation within their school district. The authors found no effect on student learning, as well as evidence that this lack of effect was due to teachers' uncertainty regarding which pedagogical practices raise student learning.

Our results show that a large-scale coaching program can be an effective policy to improve the performance of existing teachers at a reasonable cost. Rather than offering incentives for teachers to devote more time and effort to the task (something that might not be effective if teachers lack the requisite pedagogical skills), the results of this paper suggest that it is more effective to directly intervene to enhance their teaching skills.

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge (2023), "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics,* 138 (1), 1-35.

Akpur, Ugur (2020), "Critical, Reflective, Creative Thinking and Their Reflections on Academic Achievement." *Thinking Skills and Creativity,* 37, 100683.

Albornoz, Facundo, Maria Victora Anauati, Melina Furman, Mariana Luzuriaga, Maria Eugenia Podesta, and Ines Taylor (2020), "Training to Teach Science: Experimental Evidence from Argentina." *The World Bank Economic Review,* 34 (2), 393-417.

Allen, Joseph, Anne Gregory, Amori Mikami, Janetta Lun, Bridget Hamre, and Robert Pianta (2013), "Observations of Effective Teacher-Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System-Secondary." *School Psychology Review,* 42 (1), 76–98.

Angrist, Joshua, and Guido Imbens (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association,* 90 (430), 431-442.

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996), "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434), 444-455.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh (2021), "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training." *American Economic Journal: Economic Policy*, 13 (1), 36-66.

Bellés-Obrero, Cristina, and Maria Lombardi (2022), "Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru." *Economic Development and Cultural Change,* 70 (4), 1631-1669.

Bennett, Daniel, Asjad Naqvi, and Wolf-Peter Schmidt (2018), "Learning, Hygiene and Traditional Medicine." *Economic Journal,* 128 (612), F545-F574.

Bruns, Barbara, Leandro Costa and Nina Cunha (2018), "Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil?" *Economics of Education Review*, 64 (1), 214-250.

Castro, Juan, and Bruno Esposito (2022), "The Effect of Bonuses on Teacher Retention and Student Learning in Rural Schools: A Story of Spillovers." *Education, Finance and Policy*, 17 (4), 693-718.

Castro, Juan F., Paul Glewwe, Alexandra Heredia-Mayo, Stephanie Majerowicz and Ricardo Montero (2024a), "Supplement to 'Can Teaching Be Taught? Improving Teachers' Pedagogical Skills at Scale in Rural Peru'," Quantitative Economics Supplemental Material.

Castro, Juan F., Paul Glewwe, Alexandra Heredia-Mayo, Stephanie Majerowicz and Ricardo Montero (2024b), "Replication Package for: Can Teaching Be Taught? Improving Teachers' Pedagogical Skills at Scale in Rural Peru." Zenodo. https://doi.org/10.5281/zenodo.13738582.

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

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

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

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

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers (2018), "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics,* 133 (2), 993-1039.

Evans, David, and Anna Popova (2016), "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer,* 31 (2), 242-270.

Evans, David, and Fei Yuan (2022), "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis,* 44 (3), 532–540.

Fauth, Benjamin, Jasmin Decristan, Anna-Theresia Decker, Gerhard Büttner, Ilonca Hardy, Eckhard Klieme, and Mareike Kunter (2019), "The Effects of Teacher Competence on Student Outcomes in Elementary Science Education: The Mediating Role of Teaching Quality." *Teaching and Teacher Education*, 86, 102882.

Gage, Nicholas, Terrance Scott, Regina Hirn, and Ashley MacSuga-Gag (2018), "The Relationship Between Teachers' Implementation of Classroom Management Practices and Student Behavior in Elementary School." *Behavioral Disorders*, 43 (2), 302–315. https://doi.org/10.1177/0198742917714809.

Georgiadis, Andreas, and Christos Pitelis (2016), "The Impact of Employees' and Managers' Training on the Performance of Small-and Medium-Sized Enterprises: Evidence from a Randomized Natural Experiment in the UK Service Sector." *British Journal of Industrial Relations,* 54 (2), 409-421.

Jukes, Matthew, Elizabeth Turner, Margaret Dubeck, Katherine Halliday, Hellen Inyega, Sharon Wolf, Stephanie Simmons Zuilkowski, and Simon Brooker (2017), "Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial". *Journal of Research on Educational Effectiveness*, 10 (3), 449-481.

Kotze, Janeli, Brahm Fleisch, and Stephen Taylor (2019), "Alternative Forms of Early Grade Instructional Coaching: Emerging Evidence from Field Experiments in South Africa." *International Journal of Educational Development*, 66, 203-213.

Kovner, Christine, Carol Brewer, Farida Fatehi, and Jin Jun (2014), "What Does Nurse Turnover Rate Mean and What is the Rate?" *Policy, Politics, & Nursing Practice*, 15 (3-4), 64-71.

Kraft, Matthew, David Blazar, and Dylan Hogan (2018), "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research,* 88 (4), 547-588.

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

Lucas, Adrienne, Patrick McEwan, Moses Ngware and Moses Oketch. (2014), "Improving Early-grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda." *Journal of Policy Analysis and Management* 33 (4), 950-976.

Matsumura, Lindsay, Hellen Garnier, Richard Correnti, Brian Junker, and Donna DiPrima Bickel (2010), "Investigating the Effectiveness of a Comprehensive Literacy Coaching Program in Schools with High Teacher Mobility." *The Elementary School Journal,* 111 (1), 35-62.

Ministry of Education (2019a), *Administrative Files on School Characteristics, Teacher Characteristics and Student Outcomes [database].* Ministry of Education of Peru, Lima, last accessed 2019-08-01.

Ministry of Education (2019b), *Monitoring of School Practices [database].* Office of Strategic Monitoring and Evaluation. Ministry of Education of Peru, Lima, last accessed 2019-10-17.

Popova, Anna, David Evans, and Violeta Arancibia (2016), "Training Teachers on the Job: What Works and How to Measure It." Policy Research Working Paper 7834. The World Bank: Washington, DC.

Romano, Joseph, and Michael Wolf (2016), "Efficient Computation of Adjusted p-Values for Resampling-Based Stepdown Multiple Testing." *Statistics & Probability Letters,* 113, 38-40.

Schaffner, Julie, Paul Glewwe and Uttam Sharma (2024), "Why Programs Fail: Lessons for Improving Public Service Quality from a Mixed Methods Evaluation of an Unsuccessful Teacher Training Program." Forthcoming, *World Bank Economic Review*.

Stronge, James, Thomas Ward, and Leslie Gran (2011), "What Makes Good Teachers Good? A Cross-Case Analysis of the Connection Between Teacher Effectiveness and Student Achievement." *Journal of Teacher Education*, 62 (4), 339–355. https://doi.org/10.1177/0022487111404241.

Wisniewski, Benedikt, Klaus Zierer, and John Hattie (2020), "The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research." *Frontiers in Psychology*, 10, 3087.

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

Zeitlin, Andrew (2021). "Teacher Turnover in Rwanda." *Journal of African Economies*, 30 1, 81-102.