PERUVIAN ECONOMIC ASSOCIATION

Work With What You've Got: Improving Teachers' Pedagogical Skills at Scale in Rural Peru

Juan F. Castro

Paul Glewwe

Ricardo Montero

Working Paper No. 158, December 2019

Work With What You've Got: Improving Teachers' Pedagogical Skills at Scale in Rural Peru[1]

Juan F. Castro
Universidad del Pacifico

Paul Glewwe
University of Minnesota

Ricardo Montero
University of Minnesota

This version: October 2019

Abstract

We evaluate the effect of a large-scale teacher coaching program offered in a context of high teacher turnover on a broad range of pedagogical skills. Previous studies have found that small coaching programs can improve the teaching of reading and of science in a developing country setting. However, scale can compromise quality and turnover can erode compliance. It is also unclear whether general pedagogical skills can be improved through coaching. We evaluate a teacher coaching program currently serving more than 6,000 rural public schools in Peru. We exploit the random assignment of a program expansion that occurred in 2016. We find that, after two years, the program has been effective in improving pedagogical skills with an average effect between 0.24 and 0.34 standard deviations (s.d.). Accounting for non-compliance reduces the program's effect to between 0.20 and 0.30 s.d. This is below the effect found in developed countries (0.49 s.d.) but remains reasonably large considering the scale of the program and the degree of teacher turnover.

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

[1] Authors would like to thank participants at the Trade and Development Seminar in the Department of Applied Economics of the University of Minnesota for their valuable comments. Authors are also grateful to Alexandra Heredia and Hugo Fernández for excellent research assistance. Any remaining errors are ours.

1. Introduction and Motivation

Teacher quality is an essential determinant of student learning outcomes (Das et al. 2007, Clotfelter et al. 2010, Chetty et al. 2014). Nevertheless, teachers in several education systems lack mastery in the subjects they teach, specially those teaching poor students (World Bank, 2018). An important policy question is whether teacher quality can be improved while teachers remain in-service. This policy question is particularly relevant in the developing world where students in poorer areas typically get paired with considerably less skilled and less motivated teachers. In these poor contexts, in-service training offers the possibility to improve the quality of existing teachers and has the benefits of being designed and coordinated by the Ministry of Education and of receiving support from teachers' unions (Evans and Popova, 2016). Another factor that adds weight to the importance of identifying the effectiveness of teacher training programs is the important amount of resources spent by developing countries on them (Bruns and Luque, 2014). Popova et al. (2016) identify that about two thirds of the World Bank projects with educational components between 2000 and 2012 included professional development for teachers.

Evidence from the developing world regarding the effectiveness of in-service training programs is mixed, and programs vary enormously in terms of their form and content. A recent survey by Evans and Popova (2016) revealed that those programs which include face-to-face training, follow-up visits, engaging teachers to obtain their ideas, and were adapted to the local context, tend to show larger effects on learning. Coaching programs typically exhibit these features as they involve school visits, classroom observations, and the provision of personalized feedback to teachers by trained peers or coaches. As a result, coaching programs have emerged as a promising alternative to the more traditional models of in-service training based on intensive sessions offered to a large number of teachers at a centralized venue.

A recent meta-analysis conducted by Kraft et al. (2018) found that coaching programs offered in the developed world (and especially the U.S.) can produce large effects on instructional practices and student learning (with impacts of 0.49 and 0.18 standard deviations, respectively). Recent work has also demonstrated positive effects of teacher coaching in a developing country setting. Cilliers et al. (2018) compared coaching versus centralized training offered to improve the teaching of reading skills in 180

public schools in South Africa. Albornoz et al. (2018) estimated the impact of providing teachers with guidance regarding the organization, content and pedagogy of a topic (a structured curriculum) and the impact of providing this structured curriculum plus coaching to improve the teaching of science in 70 public schools in Argentina. Although results in terms of cost-effectiveness are mixed,[2] both studies found that coaching produced positive effects on learning.

Based on the evidence summarized above, three questions remain open related to the effectiveness of coaching in improving pedagogical practices in the developing world: (i) can a program implemented at scale still exhibit positive results? (ii) can teacher turnover threaten program effectiveness? and (iii) can general pedagogical skills be improved? In this paper, we address all three of these questions. To do this, we evaluate the effect of a large-scale teacher coaching program operating in a context of high teacher turnover on a broad range of pedagogical skills.

To the best of our knowledge, no previous study has evaluated the effects on pedagogy of a large-scale teacher coaching program implemented by the government in a developing country setting.[3] The majority of in-service training programs evaluated in the developing world are pilot programs or efficacy trials run by the researchers or by non-governmental organizations, and tend to be small in scale (Evans and Popova, 2016). For example, the recent studies of Cilliers et al. (2018) and Albornoz et al. (2018) involved only 180 and 70 schools, respectively. In contrast, the program evaluated in this paper has been implemented in more than 6,000 rural schools in Peru.

In their meta-analysis of 60 studies conducted in developed countries, Kraft et al. (2018) illustrated the challenges of taking coaching programs to scale by examining the relationship between the sample size and the effect of the program. They found a negative relation and highlighted the need to move beyond efficacy trials and offer more evidence on the effect of large-scale programs.

[2] On one hand, Cilliers et al. (2018) found that coaching was more cost-effective than training, with an estimated 0.57 standard deviation increase in reading proficiency per US$ 100 spent per student each year, compared to a 0.39 standard deviation increase in the case of training. On the other hand, Albornoz et al. (2018) found that coaching was cost-effective, as the unit cost per 0.1 standard deviation was more than twice the cost of using only the structured curriculum unit.

[3] Majerowicz and Montero (2018) estimate the effect of the same program evaluated in this study on learning outcomes. They find positive effects in the schools offering training (0.25-0.38 standard deviations) which persisted as long as the school retained the trained teacher. We complement these findings by focusing on pedagogical skill as a relevant mechanism linking coaching to student learning and by addressing the issues of non-compliance and selection produced by teacher turnover.

The issue of scale is relevant for the effectiveness of coaching programs because of two characteristics of this approach to in-service training. First, its success depends on the availability of qualified coaches. If these skills are scarce, increasing program coverage will likely reduce its quality. Second, classroom observation and personalized feedback requires coaches to commute between different schools. This can become very costly and can compromise program delivery if scaling-up involves serving schools located in hard-to-reach areas and lacking appropriate infrastructure and support personnel. Rural schools in developing countries typically have these characteristics, and they usually have teachers who are more in need of additional training.

Teacher turnover is another potential threat to the effectiveness of coaching programs because it reduces compliance. In fact, teachers who leave a school while the program is still being implemented can fail to receive the full "dose" of coaching. In addition, program schools receiving new teachers can end up with a staff that is only partially trained.

Teacher turnover can also generate two different intention-to-treat effects depending on whether one evaluates the pedagogical skill of the teachers who currently work in the schools offering the program or the pedagogical skill of the teachers who were in those schools when the coaching sessions were first offered. Both effects can be relevant for policy. The first effect is relevant for policymakers seeking to improve teachers' skills in a particular set of schools because, for example, those schools host disadvantaged students. Notice that this effect depends not only on the direct effect of the program on the skill of participating teachers but also on the indirect effect of the program on the composition of the skills of the teachers who leave and arrive at these schools. The second intention-to-treat effect can be relevant for policymakers seeking to improve the skills of a particular group of teachers because, for example, they have fewer skills than their colleagues.

To the best of our knowledge, only one previous study has directly addressed the issue of teacher turnover when estimating the effect of a teacher training program. Clare et al. (2010) evaluated the effect of a literacy coaching program using a sample of 32 elementary schools in Texas. They stressed how teacher turnover can represent a challenge to schools attempting to improve instruction through teacher training and estimated the effect of program participation on the reading skills of the students of the

teachers recruited to replace those who left their school during the first year of implementation. They found that teachers' program participation was associated with an improvement in the reading skills of their students. The non-random composition of their sample, however, casts doubt about the causal interpretation of their results.

Finally, it is still unclear whether general pedagogical skills can be enhanced through coaching. Most of the coaching programs evaluated in the literature focus on the pedagogical practice related to a specific topic or course. In Cilliers et al. (2018), for example, coaching focused on improving the teaching of reading skills. Kraft et al. (2018) also point out the lack of causal evidence on coaching programs for subjects other than reading and literacy. The pedagogical skills of teachers working in public schools across the developing world are, in general, poor. Thus, it remains a relevant policy question whether coaching can be a tool to improve a broad range of teachers' skills.

In this paper, we evaluate the effects of a teacher training program currently operating in more than 6,000 rural schools in Peru on teachers' pedagogical skills. These skills were measured through the observation of teacher-student interactions and comprise a broad range of instructional practices. Independent classroom observers graded from 1 (ineffective) to 4 (highly effective) the way in which teachers plan their lessons, manage class time, encourage students' critical thinking and participation, provide feedback, encourage respectful classroom relations, and manage students' behavior.

The program was launched in 2010 and was designed to serve multigrade schools located in rural areas. It consists of classroom visits carried out by trained coaches, who then provide feedback to teachers on their pedagogical practices. This feedback includes information on the specific aspects of the teacher's pedagogical practice that need to be improved, as well as customized strategies to improve them. Identification exploits the random assignment of the program expansion occurred in 2016 over a population of almost 6,200 eligible schools. Pedagogical skills were measured during the last quarter of 2017 (after almost two years of treatment).

The evaluation sample comprises a random subsample of 364 rural, multigrade schools. As in many developing countries, rural schools in Peru experience very high rates of teacher turnover. Around 43% of the teachers working in these schools in 2016 left

them by 2017. A particularly valuable aspect of the data is that classroom observations were carried out not only in the schools belonging to the evaluation sample (schools originally assigned to treatment or control), but also in many (though not all) of the schools hosting those teachers who migrated from an evaluation sample school to another school between 2016 and 2017. In other words, an effort was made to track and observe those teachers who worked in an evaluation sample school in 2016 but migrated to a school outside this sample in 2017. This design allows us to estimate the effect of offering the program for two years on the teachers who were in the program schools in year one (using the data that follow teachers who moved to different schools between early 2016 and late 2017) and the effect of offering coaching for two years on the teachers who were in the program schools in year two (using the data that follow the same schools, the evaluation sample schools, over time).

Our main findings can be summarized as follows. After two years, the program has been effective in improving teachers' pedagogical skills. The aggregate measure of pedagogical skill increased between 0.24 and 0.34 standard deviations for those teachers who received the two years of training. This improvement is concentrated on two specific dimensions: lesson planning and encouraging students' critical thinking.

Intention-to-treat estimates reveal that the effect of *offering* coaching is an increase in teachers' pedagogical skills of between 0.20 and 0.30 standard deviations. These results are less than the effect of coaching programs implemented in developed countries (0.49 standard deviations on instructional practices according to Kraft et al., 2018) but remain reasonably large considering the scale of the program and the high rate of teacher turnover.

These results confirm that turnover can erode program effectiveness, but the overall difference between the intention to treat and the treatment effect is not very large. This is because all teachers assigned to treatment received at least one year of coaching and because turnover did not directly translate into non-compliance.[4].

We did not find evidence of a statistically significant difference between the effect, after two years, of offering coaching on the skill of teachers who were working in the

[4] Around 10% of teachers changed their school between 2016 and 2017 but did not change their treatment status. Notice that the program is operating at scale and it is therefore offered in schools outside the evaluation sample.

evaluation sample schools in the first year and the effect of offering coaching on the skill of teachers who were working in these schools during the second year of the program. From the point of view of the policymaker, this means that the program is equally effective whether targeted on a group of teachers or targeted on a group of schools.

The rest of the paper is organized as follows. Section 2 describes the intervention and explains the evaluation design. Section 3 presents a simple analytical framework to clarify the differences between the two intention-to-treat effects discussed above. Section 4 presents our main results in terms of the data collection exercise and the estimated effects of the program on teachers' pedagogical skills. Finally, Section 5 closes with some concluding remarks, policy implications and avenues for further research.

2. The Coaching Program and its Evaluation Design

In 2010, the Peruvian government started to implement coaching programs to improve public teachers' pedagogical practices. Under these programs, the local education authority (UGEL), following guidelines established by the Ministry of Education, hires coaches to visit teachers in schools targeted by the program.

The work of the coaches is divided into several stages. First, they meet with the school principal and gather information about the educational context. Then, in the same visit, the coaches attend a class session of the teacher and collect information on his or her performance in the classroom to make an initial diagnostic. Based on this diagnostic, the coach identifies the competencies that the teacher must strengthen and, together with the teacher, develops a plan of improvement. After this, the coach periodically observes class sessions carried out by the teacher at regular intervals during the year. In total, nine visits are made each year. After each classroom observation, the coach and the teacher meet to discuss the progress made with respect to the improvement plan. The coach makes monthly and quarterly reports that are sent to the UGEL and to the school principal on progress and on areas for future improvement of the teacher. At the end of the year, the coach provides a final feedback session for the teacher and collects his or her impressions of the process. Finally, the coach makes a final report for each teacher

on the actions, achievements, and areas that require additional effort, with reference to the initial improvement plan.

These programs represent a substantial investment by the Peruvian government, with more than US$ 130 million being spent every year on them. By 2016, teachers in over 14,000 schools were receiving coaching through these programs, which potentially affected more than 900,000 students studying in these schools. More than 90% of the schools where the program operates are primary schools, and three versions of the coaching are offered in these schools: (i) Bilingual coaching (for schools where most of the students' native language is not Spanish but one of Peru's indigenous languages); (ii) monolingual multigrade coaching (for small schools with predominantly Spanish speaking students and where the number of teachers is less than the number of grades); and (iii) monolingual full teacher schools (for schools with enough students to justify hiring one teacher for each grade).

At the beginning of the 2016 school year, a randomization mechanism was used for the expansion of the monolingual multigrade version of the program (*Acompañamiento Pedagogico Multigrado*, in Spanish, or APM). All schools that had one or two years of treatment by the end of 2015 continued to participate in APM. Monolingual multigrade schools that had not received the program yet and had low scores in the Peruvian second grade national student evaluation were randomized into treatment and control groups. Out of 6,207 eligible schools, 3,795 schools were randomly assigned to the treatment group and started receiving APM at the beginning of the 2016 school year (the Peruvian school year runs from February to November), while the remaining 2,412 schools were sorted into the control group and did not participate in any coaching program for the following years. This randomization was stratified at the region level, which is the highest level of political division in Peru, with a total of 26 regions in the country.

A random sub-sample of 364 schools stratified at the region level was selected for this study. In particular, 182 schools were randomly selected from the 3,795 treated schools, and 182 schools were randomly selected from the 2,421 control schools. Observations of teachers' pedagogical practices were carried out in these 364 schools at the end of the 2017 school year. In addition, an effort was made to follow the teachers who worked in these 364 schools in 2016 and moved to other schools in 2017 and observations were carried out in their new schools.

3. Framework and Treatment Effects

Teacher turnover can compromise compliance. The program is a two-year intervention, yet after one year of training many teachers in the schools assigned to receive the program had moved to a school that did not offer it. In addition, many teachers originally assigned to a control group school had moved to a school that offered the program and thus received one year of treatment. From the point of view of schools, some schools assigned to offer the program might receive, in year 2 (2017), teachers with no prior training, and schools in the control group can receive teachers who have been exposed to the program in year 1 (2016).

Teacher turnover can also introduce new mechanisms though which the training program can affect teachers' pedagogical skills in the schools where it is offered. One such mechanism is that the program can affect the composition of pedagogical skill in the schools that offer it by attracting teachers with particular characteristics.

Some structure is needed to account for these two phenomena. In this section, we present the assumptions we impose so we can use the available data to estimate the effects of the training program on teachers' pedagogical skills accounting for the effects of teacher turnover.

3.1 A Production Function of Pedagogical Skill

Let us assume that pedagogical skill is a single variable that has a cumulative nature and is positively affected by experience. An additional year of experience will have a different effect depending on the teacher (some teachers take more advantage of experience than others to increase their pedagogical skills) and the school where he/she worked that year (some schools offer a coaching program).

The pedagogical skill of teacher i at the end of year t (y_{it}) is a function of three inputs: (i) the skill he/she had at the end of year -1 (y_{it-1}); (ii) the teacher-specific effect of one year of experience (λ_i); and (iii) whether the school where he/she worked during year t offered the coaching program. Assume that the presence of coaching in the school where teacher i worked during year t is identified through the indicator T_{it} ($T_{it} = 1$ if the school offered training and $T_{it} = 0$ if it did not). Thus, we can write:

$$y_{it} = F(y_{it-1}, \lambda_i, T_{it}; \theta_t) \quad (1)$$

where θ_t is a set of parameters governing the relation between y_{it} and its inputs. For simplicity we assume that there are no complementarities between the three inputs, so we can write the following linear production function:

$$y_{it} = \rho y_{it-1} + \lambda_i + \delta T_{it} \quad (2)$$

This production function indicates that, each year, teacher i carries forward a proportion ρ of the pedagogical skill previously attained (a proportion $1 - \rho$ of the skill depreciates) and accumulates a particular dose of skill from experience. In addition to this, the teacher can further enhance his or her skills by a measure of δ if he/she works in a school that offers the coaching program.

Assume that the coaching program was randomly assigned within a group of schools (henceforth, the evaluation sample) at the end of year 0 and was evaluated at the end of year 2. Therefore, the pedagogical skill of teacher i accumulated in the first two years can be expressed as:

$$\begin{aligned} y_{i2} &= \rho y_{i1} + \lambda_i + \delta T_{i2} = \rho(\rho y_{i0} + \lambda_i + \delta T_{i1}) + \lambda_i + \delta T_{i2} \\ &= \rho^2 y_{i0} + (1+\rho)\lambda_i + \delta(\rho T_{i1} + T_{i2}) \end{aligned} \quad (3)$$

It is not our objective to identify all the parameters of this production function. Thus, for simplicity, the terms $\rho^2 y_{i0} + (1+\rho)\lambda_i$ can be combined into a single teacher-specific component ($\xi_i = \rho^2 y_{i0} + (1+\rho)\lambda_i$), which implies that the pedagogical skill attained by the end of year 2 can be expressed as:

$$y_{i2} = \xi_i + \delta(\rho T_{i1} + T_{i2}) \quad (4)$$

Notice that the linearity assumption implies there is no complementarity between the teacher-specific component ξ_i and coaching. This means coaching has the same effect on every teacher.[5]

3.2 Two Intention-to-Treat Effects

[5] This assumption can and will be tested by evaluating the existence of heterogeneous treatment effects.

Teacher i can be one of the teachers who were working in an evaluation sample school during year 1 (henceforth, sample 1) or one of the teachers who were working in an evaluation sample school during year 2 (henceforth, sample 2). These two samples are not necessarily the same because teachers can change their school between years 1 and 2.

One way to estimate the effect of offering training on teachers' pedagogical skill is to use sample 1 teachers to regress the pedagogical skill measured at the end of year 2 on an intercept and these teachers' treatment status in year 1. Thus, we have:

$$y_{i2} = \alpha_1 + \beta_1 T_{i1} + \varepsilon_{1i} \tag{5}$$

The coefficient $\hat{\beta}_{1,OLS}$ estimates an intention-to-treat effect. It provides an estimate for the effect of offering training for two years on the teachers who were in the treated schools in year 1, independently of the school where they ended up working in year 2.

The coefficient $\hat{\beta}_{1,OLS}$ estimates $E[y_{i2}|T_{i1} = 1] - E[y_{i2}|T_{i1} = 0]$. According to (4), this difference in conditional means can be expressed as:

$$\begin{aligned} E[y_{i2}|T_{i1} = 1] - E[y_{i2}|T_{i1} = 0] \\ &= \big(E[\xi_i|T_{i1} = 1] + \delta(\rho + E[T_{i2}|T_{i1} = 1])\big) \\ &\quad - \big(E[\xi_i|T_{i1} = 0] + \delta(0 + E[T_{i2}|T_{i1} = 0])\big) \\ &= (E[\xi_i|T_{i1} = 1] - E[\xi_i|T_{i1} = 0]) \\ &\quad + \delta(\rho + E[T_{i2}|T_{i1} = 1] - E[T_{i2}|T_{i1} = 0]) \end{aligned} \tag{6}$$

Random assignment of training at the end of year 0 ensures $E[\xi_i|T_{i1} = 1] = E[\xi_i|T_{i1} = 0]$. In other words, the teachers who were working in treated and control schools in year 1 share similar characteristics. Thus, we have:

$$E[y_{i2}|T_{i1} = 1] - E[y_{i2}|T_{i1} = 0] = \delta(\rho + E[T_{i2}|T_{i1} = 1] - E[T_{i2}|T_{i1} = 0]) \tag{7}$$

It is reasonable to assume that there is little or no depreciation between year 1 and year 2 ($\rho = 1$) in order to focus on the consequences that teacher turnover can have on the components of $E[y_{i2}|T_{i1} = 1] - E[y_{i2}|T_{i1} = 0]$. Following (2), a single year of training

should enhance skill by a size of δ so, under perfect compliance, one can expect a direct effect of size 2δ after two years of treatment. According to (7), however, teacher turnover implies that compliance is not perfect, so that $E[y_{i2}|T_{i1}=1]-E[y_{i2}|T_{i1}=0] \neq 2\delta$.

There are two ways in which compliance is not perfect. The first is the possibility that $E[T_{i2}|T_{i1}=1]<1$, that is, the possibility that some teachers that received training in year 1 moved to a school that did not offer training in year 2. The second is the possibility that $E[T_{i2}|T_{i1}=0]>0$. This means that some teachers who worked in a school that did not offer treatment in year 1 ended up receiving training in year 2. Both reduce compliance by causing $E[T_{i2}|T_{i1}=1]-E[T_{i2}|T_{i1}=0]$ to be < 1.

Notice that, according to (7), and in the absence of depreciation, the direct effect of one year of training is given by:

$$\delta=\frac{E[y_{i2}|T_{i1}=1]-E[y_{i2}|T_{i1}=0]}{1+E[T_{i2}|T_{i1}=1]-E[T_{i2}|T_{i1}=0]} \tag{8}$$

In principle, this ratio provides the effect of one year of coaching on those teachers who were induced to take up the two-year program by being assigned to it in year 1; i.e. it is a local average treatment effect (LATE). We are assuming (and will support with evidence), however, that there is no heterogeneity in treatment effects (δ is the same for all teachers). This means that this LATE corresponds to an average treatment effect. Also notice that δ corresponds to the instrumental variable estimate of the effect of one year of coaching using its random assignment as instrument. In fact, using $\hat{\beta}_{1,OLS}$ and the sample counterparts of $E[T_{i2}|T_{i1}=1]$ and $E[T_{i2}|T_{i1}=0]$ to solve for $\hat{\delta}$ in (8), is equivalent to running y_{i2} on the number of years of coaching received, using T_{i1} as an instrument.[6]

[6] This is just an application of the Wald estimator which corresponds to an instrumental variable estimate when the instrument is a binary indicator (see, for example, Duflo, et al. 2008). The instrumental variable estimate of the effect of one round of coaching (assume teacher i had N_i rounds) on pedagogical skill (y_{i2}) using the treatment status of the school where teacher i worked in year 1 (T_{i1}) as an instrument can be expressed as: $\hat{\beta}_{IV}=\frac{E[y_{i2}|T_{i1}=1]-E[y_{i2}|T_{i1}=0]}{E[N_i|T_{i1}=1]-E[N_i|T_{i1}=0]}$. To see how this corresponds to the expression given for δ in (8), notice that N_i can be expressed as $N_i=T_{i1}+T_{i2}$, where T_{i2} is the treatment status of the school where teacher i worked in year 2. The denominator of the Wald estimator given above can, therefore, be expressed as: $E[N_i|T_{i1}=1]-E[N_i|T_{i1}=0]=E[T_{i1}+T_{i2}|T_{i1}=1]-$

Another way to estimate the effect of offering coaching on pedagogical skill is by running a regression of pedagogical skill measured at the end of year 2 on an intercept and the teachers' treatment status in year 2, using sample 2. Formally:

$$y_{i2} = \alpha_2 + \beta_2 T_{i2} + \varepsilon_{2i} \tag{9}$$

The coefficient $\hat{\beta}_{2,OLS}$ provides an estimate for $E[y_{i2}|T_{i2} = 1] - E[y_{i2}|T_{i2} = 0]$. Using (4), this difference in conditional means can be expressed as:

$$\begin{aligned} E[y_{i2}|T_{i2} = 1] - E[y_{i2}|T_{i2} = 0] \\ &= \big(E[\xi_i|T_{i2} = 1] + \delta(\rho E[T_{i1}|T_{i2} = 1] + 1)\big) \\ &- \big(E[\xi_i|T_{i2} = 0] + \delta(\rho E[T_{i1}|T_{i2} = 0] + 0)\big) \\ &= (E[\xi_i|T_{i2} = 1] - E[\xi_i|T_{i2} = 0]) \\ &+ \delta\big(1 + \rho(E[T_{i1}|T_{i2} = 1] - E[T_{i1}|T_{i2} = 0])\big) \end{aligned} \tag{10}$$

The coefficient $\hat{\beta}_{2,OLS}$ also estimates an intention-to-treat effect. It provides an estimate for the effect of assigning coaching to schools for two years, independently of the teachers who ended up working in these schools in year 2. As in the case of $\hat{\beta}_{1,OLS}$, teacher turnover can cause this effect to differ from 2δ due to imperfect compliance. In addition, teacher turnover and can introduce an indirect mechanism through which the program can affect the pedagogical skill observed in the schools that offer the program for two years.

The expression given in (10) can be used to clarify this. The first expression in parenthesis, $(E[\xi_i|T_{i2} = 1] - E[\xi_i|T_{i2} = 0])$, corresponds to the difference in the average teacher-specific component between control and treatment schools. Unlike the analogous term in equation (6), random assignment of the training program at the end of year 0 does *not* ensure that $E[\xi_i|T_{i2} = 1] = E[\xi_i|T_{i2} = 0]$. This is because the program can affect teachers' decisions to migrate to or from treated schools between years 1 and 2. For example, the program could attract more skilled teachers, in which case $E[\xi_i|T_{i2} = 1] > E[\xi_i|T_{i2} = 0]$. This composition effect is an indirect mechanism

---

$E[T_{i1} + T_{i2}|T_{i1} = 0] = 1 + E[T_{i2}|T_{i1} = 1] - E[T_{i2}|T_{i1} = 0]$, which corresponds to the denominator of the ratio given in (8).

through which training can affect the pedagogical skill in the schools offering the program.

The second expression in parenthesis on the right-hand side of (10) corresponds to the direct effect of the program on pedagogical skill. If one ignores depreciation by setting $\rho = 1$, there are two ways in which teacher turnover can erode compliance and deviate the estimated direct effect of the program from the effect of two years of training (2δ).

The first is the possibility that some teachers working in treated schools during year 2 were not exposed to this training in year 1 because they migrated from schools where the program was not implemented. This translates into $E[T_{i1}|T_{i2} = 1] < 1$. The second is the possibility that some teachers working in control schools during year 2 were exposed to the training in year 1 because they migrated at the end of year 1 from schools where the program was implemented. This means that $E[T_{i1}|T_{i2} = 0] > 0$.

Ignoring depreciation, so that ρ = 1, one can use (10) to solve for the effect of one year of coaching:

$$\delta = \frac{E[y_{i2}|T_{i2} = 1] - E[y_{i2}|T_{i2} = 0]}{(E[\xi_i|T_{i2} = 1] - E[\xi_i|T_{i2} = 0]) + (1 + E[T_{i1}|T_{i2} = 1] - E[T_{i1}|T_{i2} = 0])} \quad (11)$$

In this case, we need the additional assumption of no composition effect $(E[\xi_i|T_{i2} = 1] = E[\xi_i|T_{i2} = 0])$ for (11) to be equivalent to the instrumental variable estimate of the effect of one year of treatment, using sample 2 teachers and T_{i2} as an instrument. In fact, if one imposes $E[\xi_i|T_{i2} = 1] = E[\xi_i|T_{i2} = 0]$ and uses $\hat{\beta}_{2,OLS}$ and the sample counterparts of $E[T_{i1}|T_{i2} = 1]$ and $E[T_{i1}|T_{i2} = 0]$ to solve for δ in (10), one will obtain the estimated effect of one year of coaching on those members of the staff of the schools offering training that received the complete two-year treatment, which corresponds to the average treatment effect following the assumption of no heterogeneity in δ.

4. Results

4.1 Fieldwork Results: Attrition and Balance

The evaluation sample is comprised of 364 schools, randomly divided into 182 treated schools and 182 control schools. Fieldwork was carried out during the third quarter of

2017 and was planned in order to observe the pedagogical practices of: (i) the teachers who were working in 2016 in a school that belongs to the evaluation sample (sample 1); and (ii) the teachers who worked in 2017 in a school that belongs to the evaluation sample (sample 2). The former required visiting schools outside the evaluation sample because many sample 1 teachers changed school between 2016 and 2017.

It was not possible to observe the pedagogical practices of all the teachers belonging to sample 1 (see Table 1). In fact, attrition in sample 1 is large. This was partly due to the fact that 50 (7.6%) of the 662 sample 1 teachers left the public educational system in 2017. In addition, information on the location of teachers at the time fieldwork was planned was not up to date. According to the information on teacher location that was available at the time fieldwork was planned, the trained observers needed to visit 406 schools, including 104 outside the evaluation sample, to survey the sample 1 teachers. During fieldwork, 91.6% (372 out of 406) of these schools were visited, but in many cases the teacher could not be found because he or she was actually working in another school. Overall, as seen in Table 1, only 68.8% (455 out of 662) of the original sample 1 teachers were observed. Of the 207 sample 1 teachers who were not observed, 50 had left the teaching profession, 28 were in one of the 24 schools that were not visited, and 129 were thought to have moved to one of the schools that were visited but in fact were working in another school that was not in the planned sample of 406 schools.

Table 1
Distribution of Sample 1 Teachers and Evaluation Sample Schools

	Sample 1 teachers			Evaluation sample schools		
	Treatment	Control	Total	Treatment	Control	Total
Original	321	341	662	182	182	364
Observed	219	236	455	166	174	340
Attrition rate (%)	0.318	0.301	0.312	0.088	0.044	0.066
Difference in attrition rates	0.017 (0.036)			0.044* (0.026)		

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Turning to sample 2 teachers, as seen in Table 1, data were collected from 340 out of 364 evaluation sample schools, and we were able to collect information from 640 teachers (341 in control schools and 299 in treated schools). It is not possible to calculate an exact attrition rate at the teacher level for sample 2 because we do not know the number of teachers that worked in the 24 schools we were unable to visit. It is reasonable to assume, however, that this number is small because unobserved schools represent only 6.6% of the sample of schools, and teachers are fairly evenly distributed across schools.

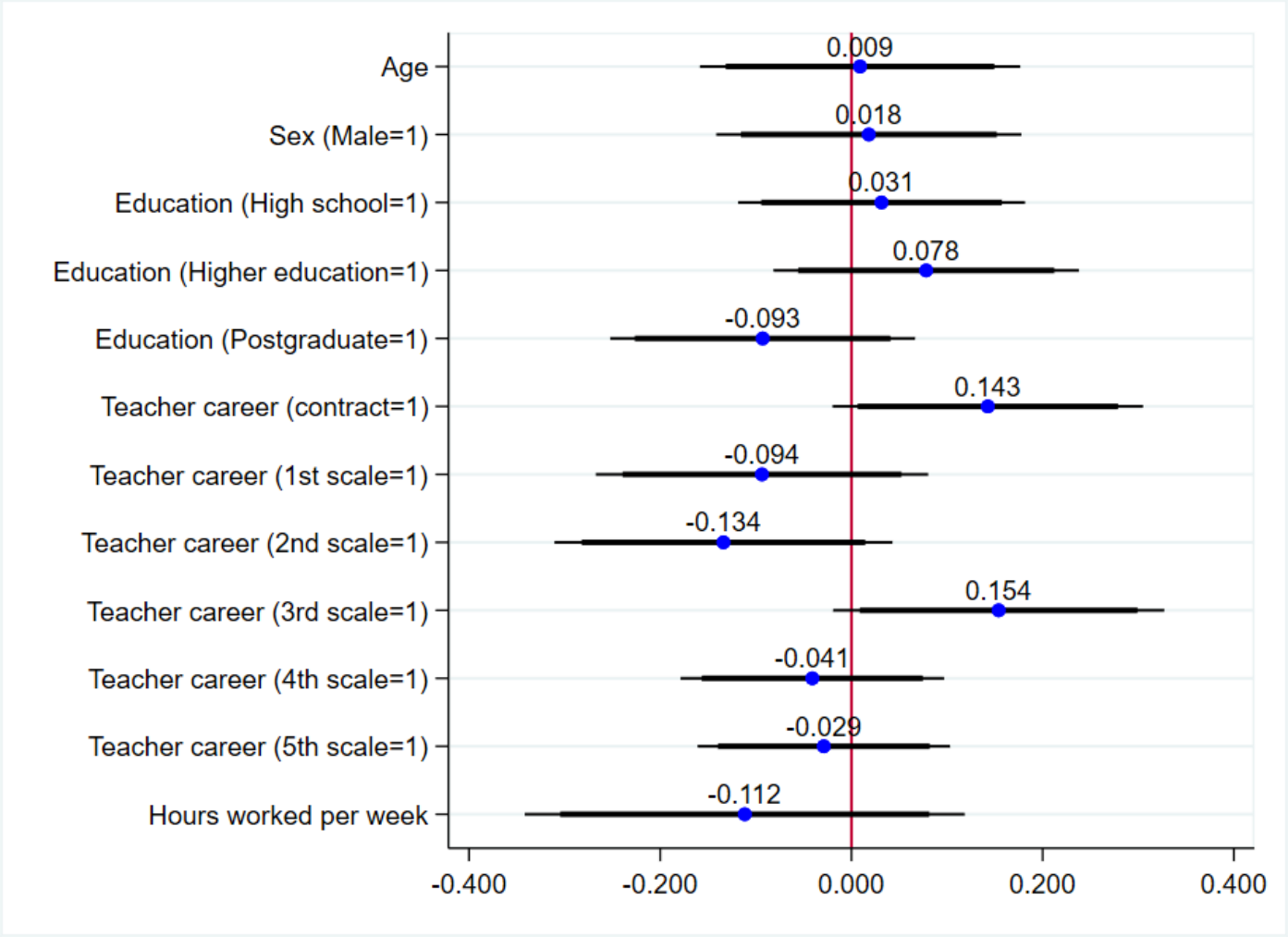
Non-random attrition could lead to biased estimates (especially estimates using sample 1). However, if the missing teachers have not affected the average characteristics of observed control and treatment teachers in different ways, then attrition will not yield biased estimates. To check for the possibility of bias, we compare observable characteristics of schools and teachers belonging to the control and treatment groups.

Random assignment of the program in 2016 should ensure that teacher characteristics were balanced for the teachers working in the control and treatment schools in that year (sample 1 teachers), that is before any attrition occurred. If attrition has not introduced a bias, we should also observe that teacher characteristics are similar between those working in control and treatment schools in the observed subsample of sample 1 teachers (the 455 teachers in Table 1). Random assignment should also ensure that the characteristics of the schools belonging to the evaluation sample are balanced between control and treatment schools.

As noted in the previous section, random assignment will not ensure that teacher characteristics are balanced between those working in control and treatment schools in 2017 (sample 2). In fact, significant correlation of these characteristics with the treatment status of the school would be evidence that the program has affected the composition of teacher characteristics in year 2. This will be tested in the next section.

Figures 1 through 4 provide evidence that the control and treatment groups share similar characteristics in terms of: (i) teacher characteristics in the original 662 sample 1 teachers; (ii) teacher characteristics in the subsample of 455 sample 1 teachers that were observed in year 2 (2017); (iii) school characteristics in the original 364 evaluation sample schools; and (iv) school characteristics in the subsample of 340 schools that were visited in year 2. More specifically, none of the (standardized) differences is very large, and none is statistically significant even at the 10% level.

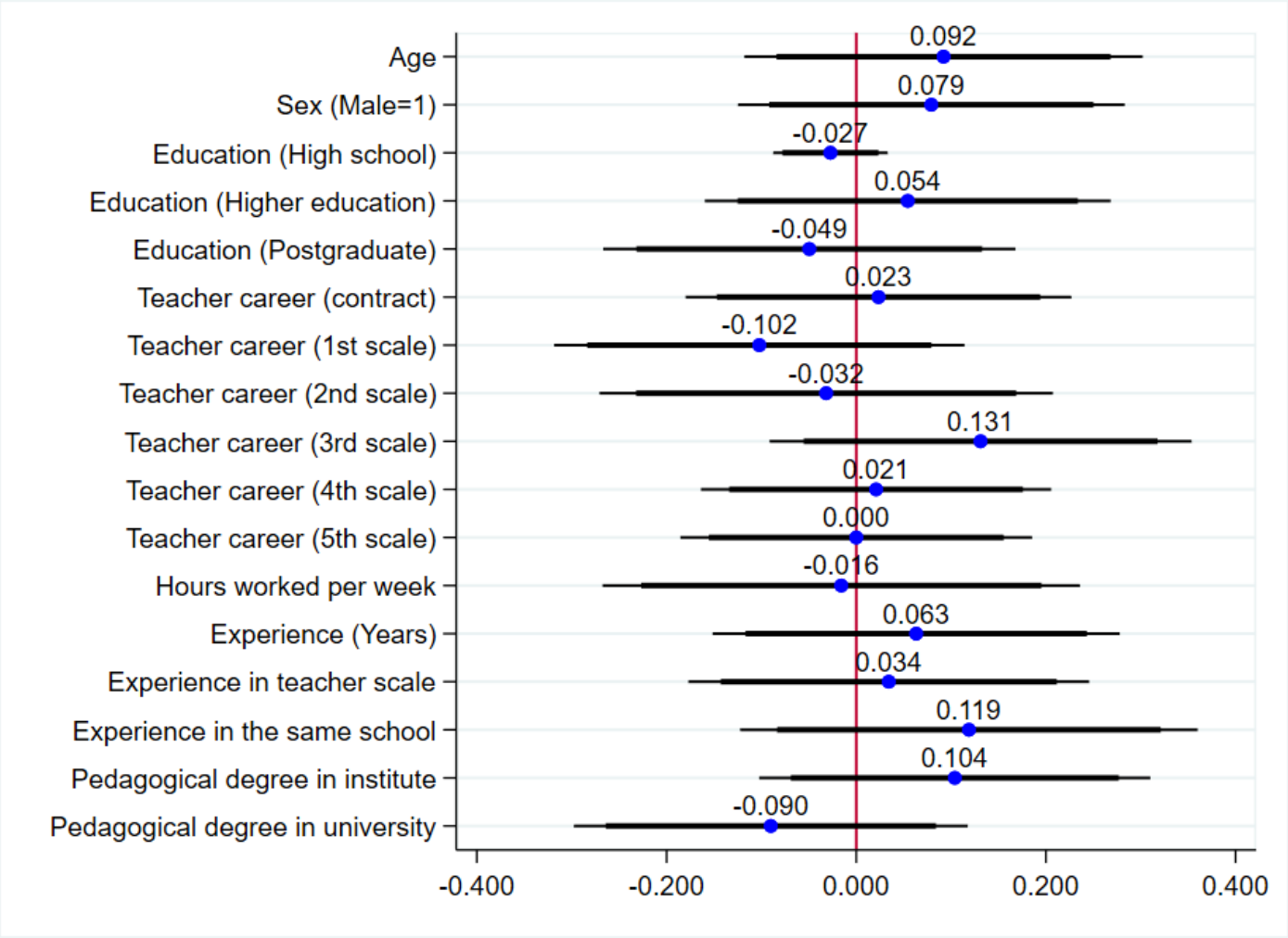
Figure 1
Balance in Teacher Characteristics for the Original 662 Teachers Who Worked in an Evaluation Sample School in 2016 (sample 1)



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.

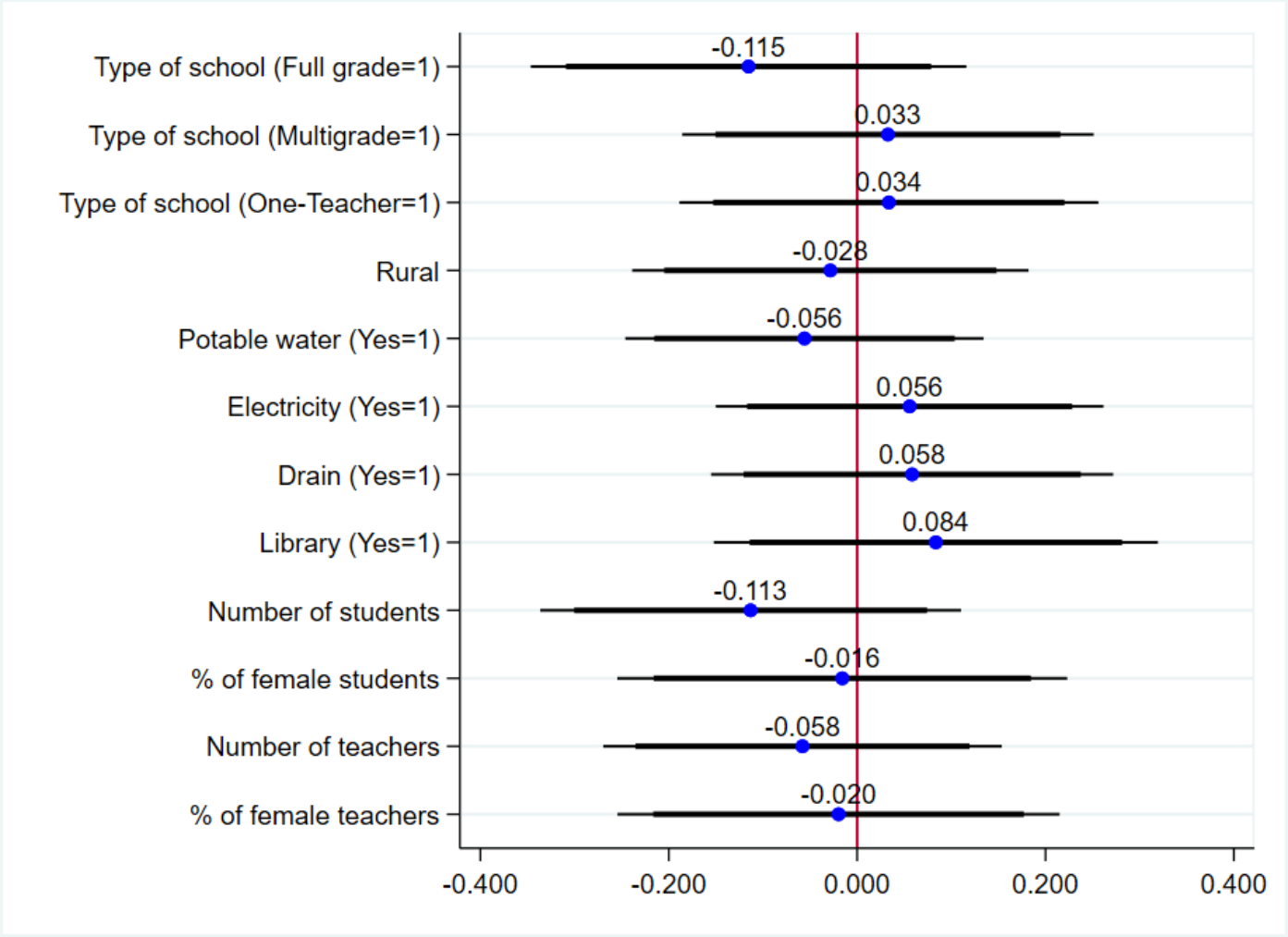
Figure 2
Balance in Teacher Characteristics for the 455 Teachers Observed in Year 2 Who Worked in an Evaluation Sample School in Year 1 (sample 1)



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.

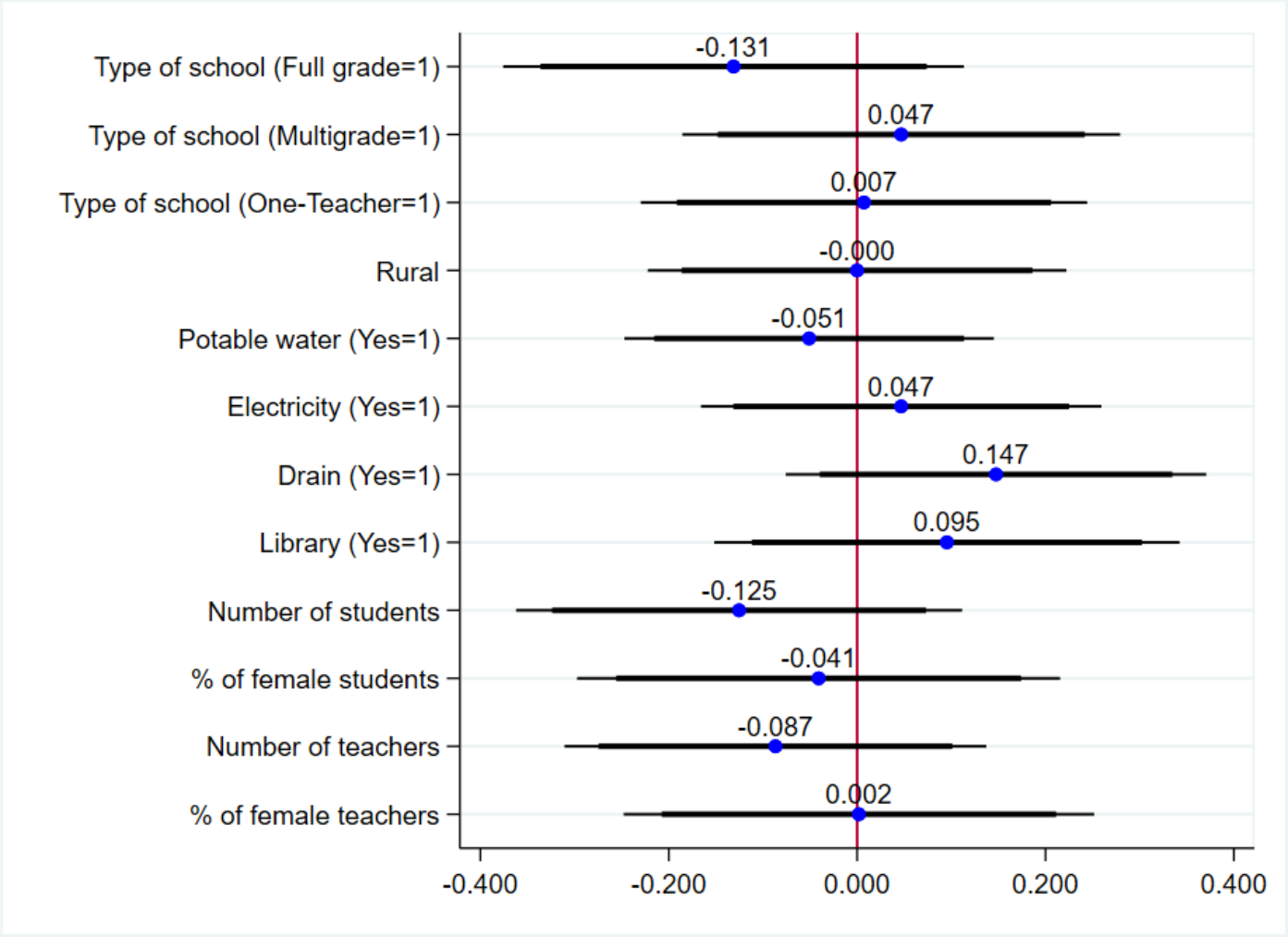
Figure 3
Balance in School Characteristics in the Original 364 Evaluation Sample Schools



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.

Figure 4
Balance in School Characteristics in the Subsample of 340 Evaluation Sample Schools that Were Visited



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.

4.2 Intention-to-Treat Estimates

In this section we present estimates for $E[y_{i2}|T_{i1}=1]-E[y_{i2}|T_{i1}=0]$ using sample 1 and $E[y_{i2}|T_{i2}=1]-E[y_{i2}|T_{i2}=0]$ using sample 2. This was done using a single index of pedagogical practice (y_{i2}) averaging the standardized scores of the eight indicators obtained during the classroom observations.

The baseline specifications to estimate these differences in outcomes are given in equations (5) (for sample 1) and (9) (for sample 2). We also include teacher characteristics as covariates when using sample 1.[7] The results are presented in Table 2. Columns (1) and (2) show the results obtained using sample 1, Column (3) displays the results for sample 2. Columns (1) and (2) show that impact of offering the program for two years on the teachers who were in the program schools in year one is an increase of approximately 0.3 standard deviations on their aggregate pedagogical skill. This result is robust to the inclusion of teacher characteristics as covariates. Column (3) shows that the impact of offering the program for two years on the teachers who were in the program schools in year two is an increase of 0.2 standard deviations on the aggregate pedagogical skill of those teachers.[8]

[7] The use of teacher characteristics as covariates is appropriate only for sample 1 because teacher characteristics observed in sample 2 can be affected by treatment. In Table A.1 in Appendix 1, we test for interactions when estimating the intention to treat effect on the pedagogical skill of sample 1 teachers. The results indicate that there is no heterogeneity by teacher experience, type of contract, position in the teacher career or sex. These results are important as they provide evidence to support the linearity assumption imposed in the production function presented in Section 2.

[8] Point estimates are somewhat smaller if we include observer fixed effects, but the general conclusions of this section remain unchanged. We present these estimates in Table A.2 in Appendix 1.

Table 2
Aggregate Skill: Intention-to-Treat Estimates

	Sample 1		Sample 2
	(1)	(2)	(3)
Treatment	**0.287*** **	**0.314*** **	**0.195** **
	(0.108)	**(0.102)**	**(0.097)**
Experience	--	0.000	--
		(0.009)	
Contract Teacher	--	0.152	--
		(0.162)	
Magisterial Level	--	0.114**	--
		(0.046)	
Sex (Men=1)	--	-0.313***	--
		(0.099)	
Age	--	-0.029***	--
		(0.009)	
R^2	0.29	0.37	0.23
N	455	455	640

* p<0.1; ** p<0.05; *** p<0.01

Note: All regressions include UGEL fixed effects. Standard errors clustered at the school level are reported in parenthesis.

4.3 Instrumental Variable Estimates of Treatment Effects

It is possible to estimate the average effect of one year of treatment by instrumenting the number of years of coaching received by the teacher using his/her treatment status (T_{i1} for sample 1 teachers and T_{i2} for sample 2 teachers). Table 3 presents the results of this instrumental variable strategy. Columns (1) and (2) show the results for sample 1, Column (3) presents the estimates for sample 2. Recall from the discussion in Section 3 that one needs the assumption of no composition effect in order to interpret the instrumental variable estimate reported Column (3) as the effect of one round of training. We will provisionally make this assumption here and explore the possibility of a composition effect in the next subsection.

Column (2) shows that one year of training increases by 0.17 standard deviations the pedagogical skill of sample 1 teachers who participated in the program. Column (3) shows that it increased by 0.12 standard deviations the pedagogical skill of sample 2 teachers who received the training. Although these point estimates are somewhat different, is not possible to reject the null hypothesis of equal treatment effects in both

samples $\left(\delta_{sample1} = \delta_{sample2} = \delta\right)$ which is consistent with the assumption that one year of treatment has the same effect on every teacher. As expected, these estimates are somewhat larger than (half of) those in Table 2 as the ITT estimates are reduced by imperfect compliance.

Table 3
Aggregate Skill: Instrumental Variable Estimates

	Sample 1		**Sample 2**
	(1)	(2)	(3)
Intensity (Years of APM)	**0.159*** **	**0.174*** **	**0.122** **
	(0.054)	**(0.050)**	**(0.056)**
Experience	--	-0.000	--
		(0.008)	
Contract Teacher	--	0.145	--
		(0.145)	
Magisterial Level	--	0.113***	--
		(0.041)	
Sex (Men=1)	--	-0.315***	--
		(0.089)	
Age	--	-0.028***	--
		(0.008)	
R^2	0.29	0.37	0.23
N	455	455	640

* p<0.1; ** p<0.05; *** p<0.01

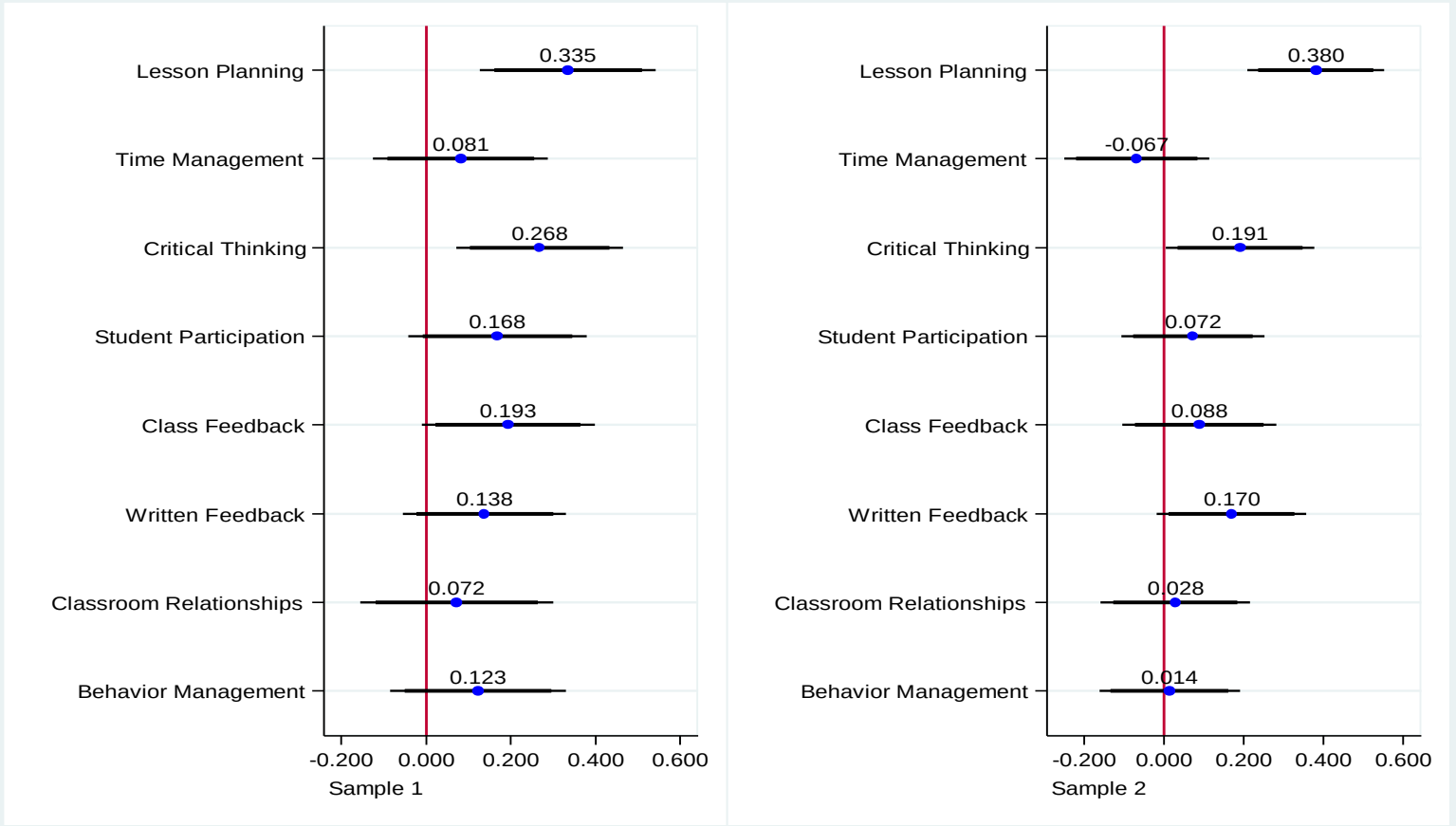
Note: All regressions include UGEL fixed effects. Standard errors clustered at the school level are reported in parenthesis.

We also estimated the effects of the program over each of the 8 pedagogical practices that contribute to the aggregated index. Figures 5 and 6 show the ITT and IV estimations, respectively. We find strong evidence that the program improves lesson planning done by the teachers (ITT of 0.335 standard deviations for sample 1 and 0.380 for sample 2, and IV of 0.186 standard deviations per year for sample 1 and 0.235 for sample 2) as well as the promotion of critical thinking (ITT of 0.268 standard deviations for sample 1 and 0.191 for sample 2, IV of 0.148 per year for sample 1 and 0.118 for sample 2).

We also find evidence that the program improves in class (oral) feedback and written feedback, but these results are less robust. We find no evidence of positive or negative

impacts of the program on any of the other four pedagogical skills (time management, student participation, classroom relationships and behavior management).

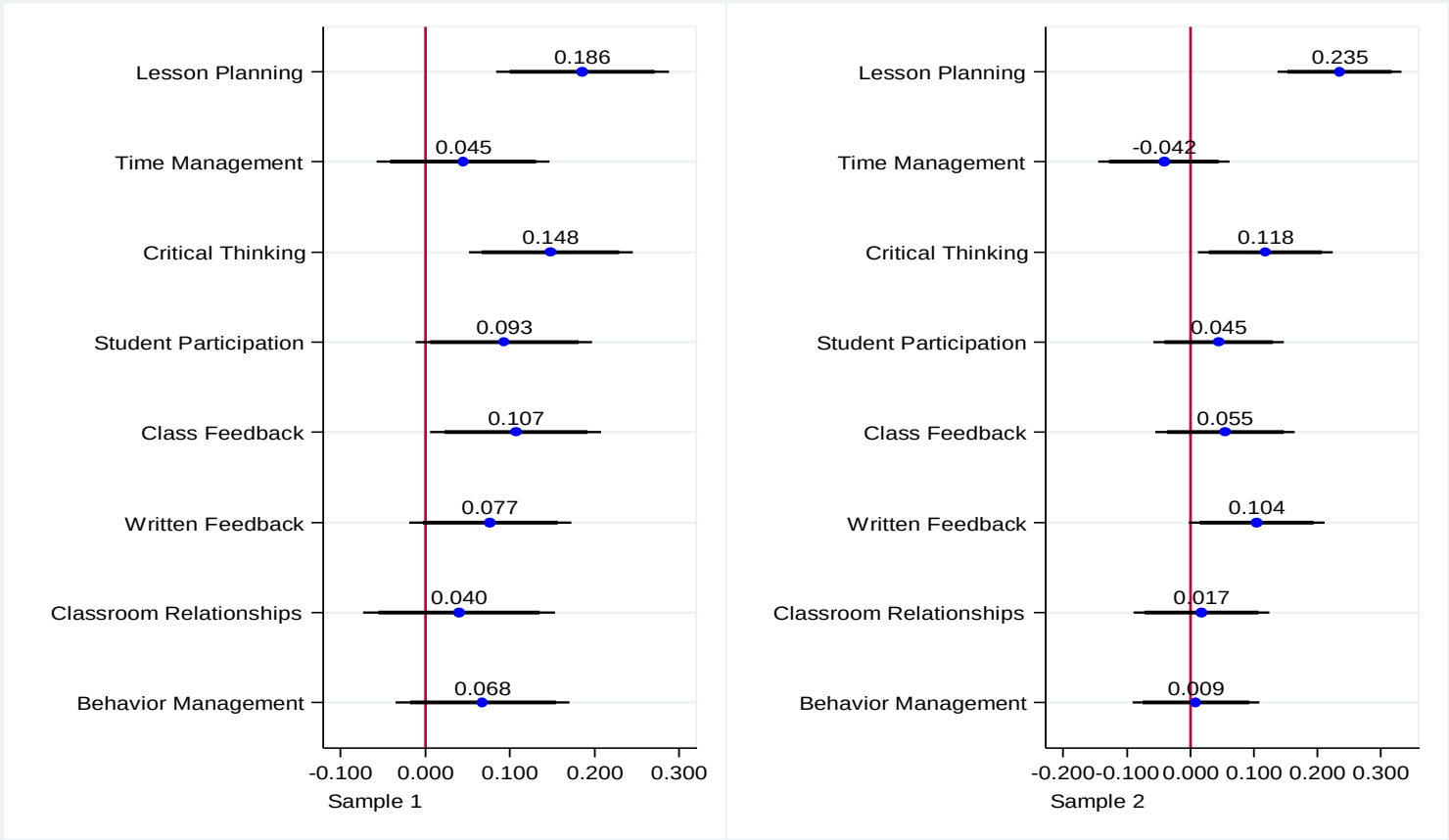
Figure 5
Disaggregated Skills: Intention to Treat Estimates



Effects measured in standard deviations. 90% and 95% confidence intervals shown

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

Figure 6
Disaggregated Skills: Instrumental Variable Estimates



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

4.4 Teacher Turnover: Compliance and Composition Effect

The framework presented in Section 3 revealed that teacher turnover can produce two different intention-to-treat estimates: the effect of training on those teachers who were working in an evaluation sample school during year 1, and the effect of training on those teachers who were working in an evaluation sample school during year 2. As argued in the Introduction, these two effects can be relevant for policy. The first will be relevant for a policymaker concerned about the skills of a certain group of teachers. The second effect will be relevant for a policymaker concerned about the pedagogical skill of teachers in a certain group of schools.

Results presented in Table 2 show that the point estimates obtained for these two intention-to-treat effects differ by around 0.1 standard deviations, although this difference is not statistically significant. According to expressions (8) and (11), one can obtain different intention-to-treat effects (in the numerator of these expressions) because of differences in compliance ($E[T_{i2}|T_{i1}=1]-E[T_{i2}|T_{i1}=0]$ for sample 1 is not the same as $E[T_{i1}|T_{i2}=1]-E[T_{i1}|T_{i2}=0]$ for sample 2) and because of the composition effect affecting sample 2 ($E[\xi_i|T_{i2}=1]-E[\xi_i|T_{i2}=0]\neq 0$).

In what follows we provide estimates for the compliance rates in both samples and for the composition effect. Compliance rates can be directly estimated from the data. For this, Tables 4 and 5 report the distribution of sample 1 and sample 2 teachers according to their destination and origin, respectively. Destinations for sample 1 teachers are classified into three categories: (i) the same school where the teacher worked in 2016; (ii) a school different from the one where the teacher worked in 2016 that offers the training program (exposed to APM); and (iii) a school different from the one where the teacher worked in 2016 that does not offer the training program (not exposed to APM).

Based on this classification, one can estimate compliance ($E[T_{i2}|T_{i1}=1]-E[T_{i2}|T_{i1}=0]$) using the proportion of treated teachers who remained in their same school or migrated to a school that offered the program in 2017, minus the proportion of control teachers who migrated to a school that offered the program in 2017. According to the estimates provided in Table 4, compliance was 82.2% for the observed sample 1 teachers.

Table 4
Distribution of Observed Sample 1 Teachers According to Their Destination School

2016 school / Destination	Treated		Control	
	No.	%	No.	%
Same School	179	0.818/a	200	0.848
Exposed to APM	13	0.059/b	13	0.055/c
Not exposed APM	27	0.123	23	0.097
	219	**1.00**	**236**	**1.00**

Note: For simple 1 teachers, /a + /b = 0.877 estimates $E[T_{i2}|T_{i1} = 1]$, /c = 0.055 estimates $E[T_{i2}|T_{i1} = 0]$, and compliance = /a + /b - /c = 0.822.

Now consider the sample 2 teachers. Table 5 shows the distribution of those teachers according to their origin. The categories are the same as those considered for the 2017 destination of sample 1 teachers. Teachers working in an evaluation sample school in 2017 can come from their same school, from a different school offering the program (exposed to APM), or from a different school not offering the program (and thus not exposed to APM).

Based on these classifications, we can estimate compliance ($E[T_{i1}|T_{i2} = 1] - E[T_{i1}|T_{i2} = 0]$) considering the proportion of treated teachers that, in 2016, worked in their same school or worked in a different school offering the program, minus the proportion of control teachers that, in 2016, worked in a school offering the program. The results, presented in Table 5, indicate that compliance in sample 2 is 61.6%.

Table 5
Distribution of Sample 2 Teachers According to Their School of Origin

2017 school / Origin	Treated		Control	
	No.	%	No.	%
Same school	179	0.599/a	200	0.587
Exposed to APM	34	0.114/b	33	0.097/c
Not exposed to APM	86	0.287	108	0.316
	299	**1.00**	**341**	**1.00**

For sample 2 teachers, /a+/b = 0.713 estimates $E[T_{i1}|T_{i2} = 1]$, /c = 0.097 estimates $E[T_{i1}|T_{i2} = 0]$, and compliance = /a + /b - /c = 0.616.

To estimate the composition effect, one can solve for the term $E[\xi_i|T_{i2} = 1] - E[\xi_i|T_{i2} = 0]$ in equation (11) replacing δ with the IV estimate of the effect of one year of training obtained for sample 1, $E[y_{i2}|T_{i2} = 1] - E[y_{i2}|T_{i2} = 0]$ with the intention-to-treat estimate obtained using sample 2, and $1 + E[T_{i1}|T_{i2} = 1] - E[T_{i1}|T_{i2} = 0]$ with

the compliance rate estimated using the transitions of sample 2 teachers reported in Table 5. The assumption is that the effect of one year of training is the same for all teachers. In Table 6 we present the results of these estimations based on bootstrapping samples 1 and 2 (the empirical distributions of the four estimates are presented in Appendix 2). We rely on repeated sampling to be able to assess the statistical significance of the estimated composition effect.

Table 6
Composition Effect

Parameter	**Estimate**/1
Effect of one round of treatment in sample 1: δ	0.162*** (0.0519) [0.0028]
Intention-to-treat effect in sample 2: $E[y_{i2}\|T_{i2}=1]-E[y_{i2}\|T_{i2}=0]$	0.201** (0.0910) [0.028]
Compliance rate in sample 2: $E[T_{i1}\|T_{i2}=1]-E[T_{i1}\|T_{i2}=0]$	0.614*** (0.0225) [0.000]
Implicit composition effect: /2 $E[\xi_i\|T_{i2}=1]-E[\xi_i\|T_{i2}=0]$	-0.371 (0.5998) [0.362]

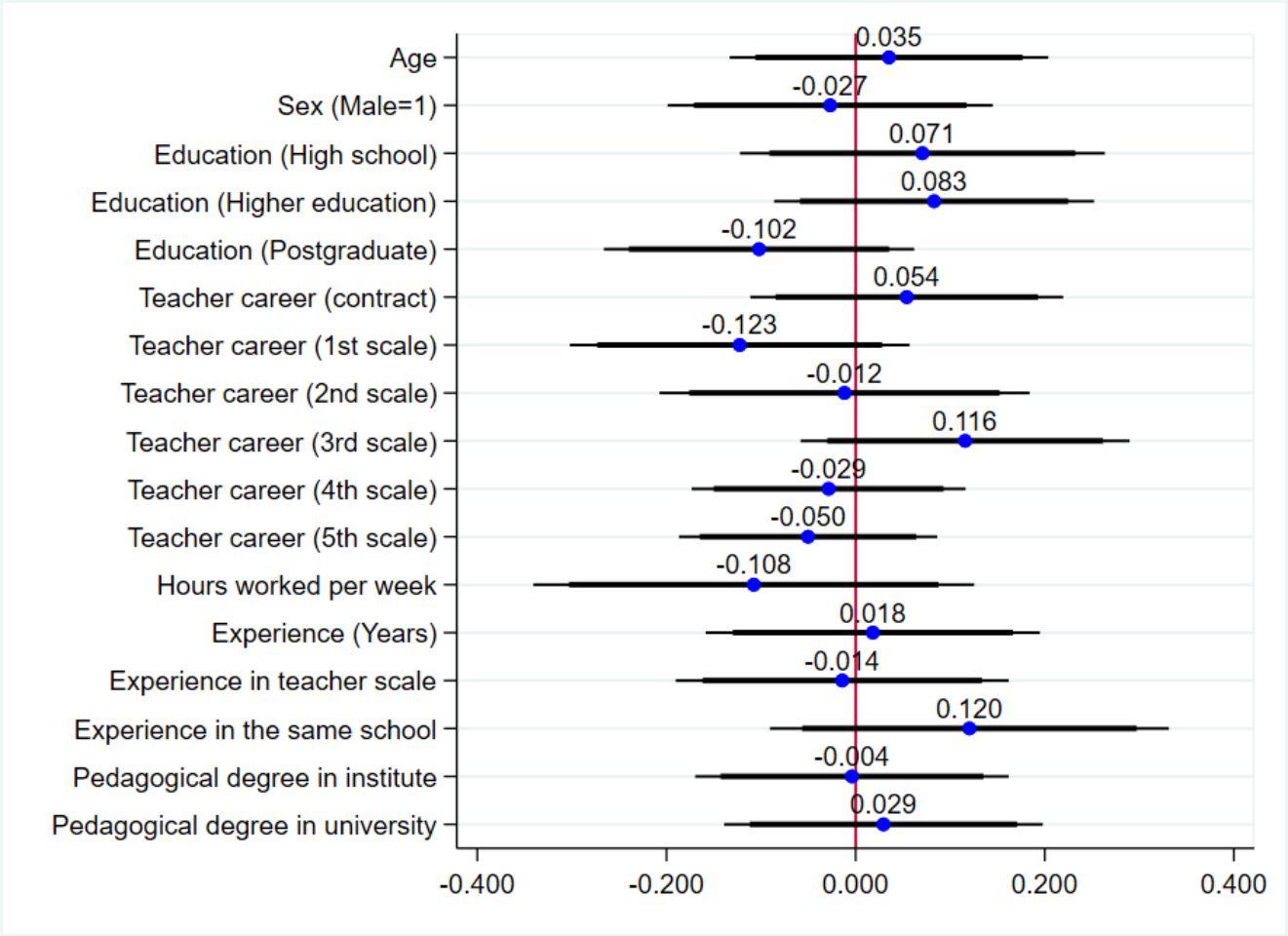
/1 Based on 5,000 repetitions of sample 1 and sample 2. Bootstrapped standard errors in parentheses and p-value for the null hypothesis of 0 effect in brackets.
P-value for 0 calculated using: $pvalue = \frac{2}{5000} min\{\#estimates < 0, \#estimates > 0\}$.
/2 We solve for the composition effect using: $E(\xi_i|T_{i2}=1)-E(\xi_i|T_{i2}=0) = [E(y_{i2}|T_{i2}=1)-E(y_{i2}|T_{i2}=0)]/\delta - [1+E(T_{i1}|T_{i2}=1)-E(T_{i1}|T_{i2}=0)]$.

Bootstrapped averages obtained for the effect of one round of treatment in sample 1, the intention-to-treat effect in sample 2, and the compliance rate in sample 2 are very similar to the point estimates obtained with the original samples (see Tables 3, 2 and 5), are and also highly significant. The estimated composition effect is -0.37 standard deviations, but we cannot reject the null that the effect is equal to 0 at standard significance levels (its standard deviation is 0.6 and the p-value for the null hypothesis of 0 effect in the distribution of estimated composition effects is 0.36).

Finally, consider an additional piece of evidence about the existence of a composition effect. Figure 5 explores the correlation between the characteristics of sample 2 teachers and the treatment status of the school where they worked in 2017. If there is a significant composition effect, one can expect a significant correlation consistent with the fact that the average teacher-specific component is different between control and treatment schools. Figure 5 shows no evidence of such correlation.

Based on the results discussed above, it is not possible to find evidence of a significant composition effect in the impact that the two-year coaching program had on the pedagogical skill of the teachers working in the evaluation sample schools in year 2. This suggests that the small difference encountered in the intention-to-treat estimates obtained with samples 1 and 2 are due to the differences in compliance rates. This also suggests that one can interpret the IV estimates based on sample 2 as the effect of one round of training and attribute the small difference with respect to the point estimate obtained from sample 1 to sampling error.

Figure 5
Treatment Effects on the Composition of Teacher Characteristics among the Teachers Who Worked in Evaluation Sample Schools in 2017 (sample 2)



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.

5. Concluding remarks

We 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 practices. We found that, after two years, the program has been effective in improving teachers' pedagogical skills with an average effect between 0.24 and 0.34 standard deviations. This effect concentrated on two dimensions of the pedagogical practice: lesson planning and encouraging students' critical thinking.

We confirmed that teacher turnover erodes compliance and reduces program effectiveness but the differences between intention to treat and treatment effects are not large. In fact, we found that the effect of *offering* coaching was between 0.20 and 0.30 standard deviations.

This analysis 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 general pedagogical skills can be improved. Moreover, we explored the issue of turnover by developing an analytical framework that explained the differences between the intention-to-treat effect for teachers who had been working in the evaluation sample schools in the first year and the intention-to-treat effect for teachers who were working in these schools in the second year of the program. According to the framework, these differences can be caused by a discrepancy in the compliance rates of both groups of teachers or by a shift in the composition of pedagogical skill in treated schools (treated schools can attract or repel teachers with particular characteristics). Although we could not find evidence of a significant difference between the two intention-to-treat effects or evidence of a significant composition effect, we believe that this framework can be useful for future evaluations carried out in contexts of high teacher turnover.

The fact that we could not find a significant difference between the two intention-to-treat effects means that assigning APM appears to be equally effective in improving the skill of the teachers originally working in the targeted schools as it is in improving the skill of the teachers who were working in these schools in year 2. From the point of view of the policymaker, this means that the program is equally effective if targeted on a group of teachers or targeted on a group of schools.

This research also contributes to the discussion about which is the most cost-effective way to improve the pedagogical skill of teachers serving rural schools and improve the performance of incumbent teachers. Rural schools typically host disadvantaged students who are in need of especially talented instructors. Rural schools are also located in hard-to-reach areas which tend to be avoided by teachers if given the choice. One way to improve pedagogical skills and student learning in rural schools is by offering incentives to attract more talented teachers. The rural bonus scheme in Peru pursues this objective by offering an approximate 30% salary increase to those teachers who take a placement in a rural school. This bonus has had a small effect on the probability of filling a teacher vacancy but has shown no effects on learning outcomes (Castro and Esposito, 2018).

The cost of the coaching program evaluated in this study is round US$ 3,000 per teacher, per year. This amount represents approximately 30% of the average annual salary of a primary education teacher in Peru. This figure is similar to the wage premium offered by the bonus program with two important differences: coaching is only a two-year investment (not a permanent salary rise) and it has produced positive results on the performance of teachers.

Developing countries with a long history of poor learning outcomes have a large mass of incumbent public teachers with poor performance. Efforts to increase the productivity of these teachers usually put a large pressure on the budget of the education sector. The literature has shown that expensive policies based on large unconditional salary rises can reduce the number of teachers taking second jobs but have no effects on the productivity of incumbent teachers (de Ree et al., 2018).

Pay-for-performance programs also offer an alternative to improve teachers' productivity. The impact of this type of incentives has been examined in several low and middle-income countries with mixed results. Very few studies, however, have estimated the effect of these programs in the context of a national intervention. One recent study evaluated the effect of a national pay-for-performance program implemented in 2015 in public secondary schools in Peru (see Bellés-Obrero and Lombardi, 2019). The program is called *Bono Escuela* and 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. Bellés-Obrero and

Lombardi (2019) found no effect on student learning and evidence that this lack of effect can be related to teachers' unawareness regarding which pedagogical practices lead to better scores.

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 incumbent teachers to devote more time and effort to the task (something which might not be effective if teachers lack the pedagogical skill), this paper shows that it is more effective to directly intervene to enhance their teaching skills.

References

Albornoz, F., Anauati, M., Furman, M., Luzuriaga, M., Podesta, M., & Taylor, I. (2018). Training to teach science: experimental evidence from Argentina. *The World Bank*.

Belles-Obrero, C., & Lombardi, M. (2019). Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru. *IZA Discussion Paper No. 12600*.

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

Castro, J., & Esposito, B. (2019). The Effect of Bonuses on Teacher Behavior: A Story with Spillovers. *Peruvian Economic Association Working Paper No 104*.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2019). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*.

Clare, L., Garnier, H., Junker, B., & Correnti, R. (2010). Investigating the Effectiveness of a Comprehensive Literacy Coaching Program in Schools with High Teacher Mobility. *The Elementary School Journal, 111*(1), 35-62.

Clotfelter, C., Ladd, H., & Vigdor, J. (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, J., Dercon, S., Habyarimana, J., & Krishnan, P. (2007). Teacher Shocks and Student Learning. Evidence from Zambia. *Journal of Human Resources, 42*(4), 820-862.

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2018). Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia. *Quarterly Journal of Economics, 133*(2), 993-1039.

Duflo, E., Glennerster, R., & Kremer, M. (2008). Using Randomization in Development Economics Research: A Toolkit. (T. Schultz, & J. Strauss, Edits.) *Handbook of Development Economics, 4*, 3895-3962.

Evans, D., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *The World Bank Economic Review, 31*(2), 242-270.

Kraft, M., Blazar, D., & Hogan, D. (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.

Majerowicz, S., & Montero, R. (2018). Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru. *Working Paper*.

Popova, A., Evans, D., & Arancibia, V. (2016). Training Teachers on the Job: What Works and How to Measure It. *Policy Research Working Paper 7834, World Bank.*

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

Appendix 1

Table A.1
Heterogeneous Treatment Effects in Sample 1

	(1)	(2)	(3)	(4)	(5)
Treatment	0.314***	0.213	0.314***	0.273*	0.236
	(0.102)	(0.240)	(0.117)	(0.159)	(0.147)
Experience	0.000	-0.002	0.000	0.000	0.000
	(0.009)	(0.011)	(0.009)	(0.009)	(0.009)
Contract Teacher	0.152	0.153	0.154	0.154	0.140
	(0.162)	(0.163)	(0.226)	(0.163)	(0.162)
Magisterial Level	0.114**	0.115**	0.114**	0.102*	0.114**
	(0.046)	(0.046)	(0.046)	(0.060)	(0.046)
Sex (Men=1)	-0.313***	-0.315***	-0.313***	-0.313***	-0.396***
	(0.099)	(0.099)	(0.099)	(0.099)	(0.147)
Age	-0.029***	-0.029***	-0.029***	-0.029***	-0.029***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Treatment #Experience		0.005			
		(0.011)			
Treatment #Contract			-0.004		
			(0.247)		
Treatment #M. Level				0.025	
				(0.081)	
Treatment #Sex					0.170
					(0.188)
R^2	0.37	0.37	0.37	0.37	0.37
N	455	455	455	455	455

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Note: All regressions include UGEL fixed effects. Standard errors clustered at the school level are reported in parenthesis.

Table A.2
Intention-to-Treat Effects with Observer Fixed Effects

	Sample 1		**Sample 2**
	(1)	(2)	(3)
Treatment	**0.215***	**0.261****	**0.162***
	(0.110)	**(0.104)**	**(0.097)**
Experience		0.003	
		(0.010)	
Contract Teacher		0.180	
		(0.168)	
Magisterial Level		0.119**	
		(0.048)	
Sex (Men=1)		-0.331***	
		(0.096)	
Age		-0.031***	
		(0.009)	
R^2	0.29	0.37	0.31
N	455	455	640

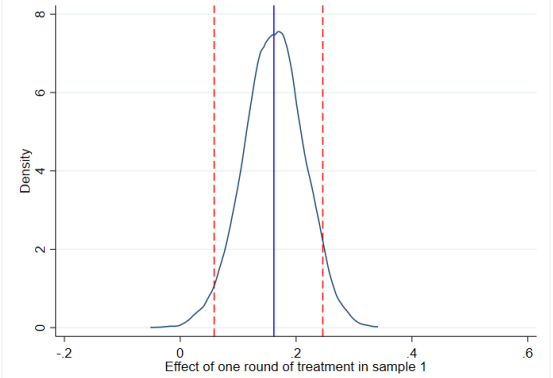
* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Note: All regressions include UGEL and observer fixed effects.
Standard errors clustered at the school level are reported in parenthesis.

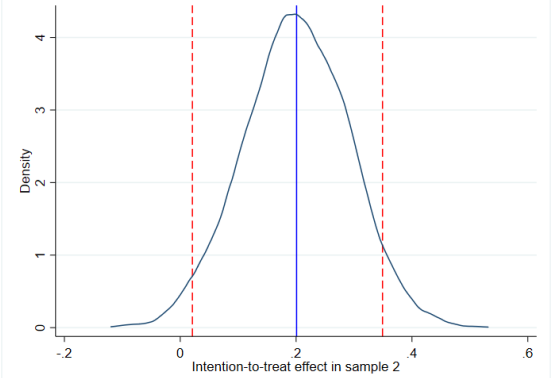
Appendix 2

Empirical distributions after 5,000 replications of sample 1 and sample 2

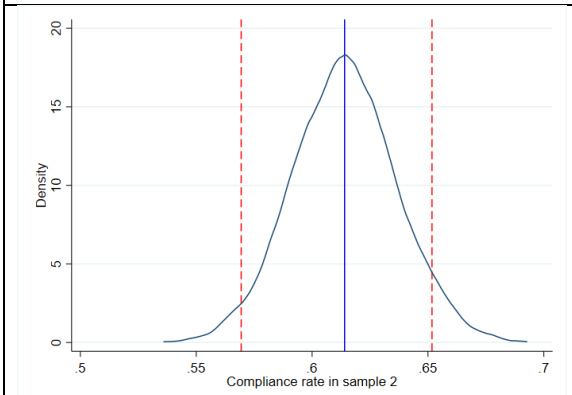
(A) Effect of one round of treatment in sample 1



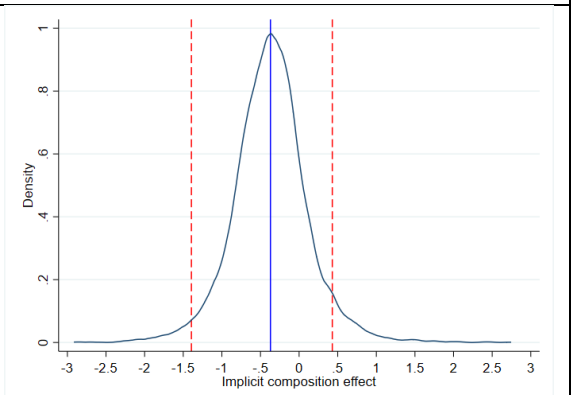
(B) Intention-to-treat effect in sample 2



(C) Compliance in sample 2



(D) Implicit composition effect



Note: Blue lines indicate the mean of the empirical distribution. Red lines indicate the 5th and 95th percentiles of the empirical distribution.