# Getting Teachers Back to School: Teacher Incentives and Student Outcomes*

Patricio Araya-Córdova† Dante Contreras‡ Jorge Rodríguez§ Paulina Sepúlveda¶

August 2022

**Abstract**

This paper studies the effects of a large-scale pay-for-performance program that attached teacher pay with measures of teaching skills. Using a unique administrative data set of over 500,000 student-teacher matches, we estimate the effect of the policy on student performance exploiting the program's gradual roll-out in a DiD analysis. We document precise, null effects on student math and language standardized test scores. Estimating a structural model of teacher skills, effort choices, and student performance, we show that making the program's financial incentives more homogeneous across teachers can improve the policy's effects on student outcomes.

*We thank Derek Neal, Michela Tincani, Sergio Urzúa, and seminar participants at NASMES 2021, SECHI 2021, XII Ridge Forum, LACEA-LAMES, and FEN-Universidad de Chile for valuable comments. Contreras acknowledges the financial support provided by COES ANID/FONDAP/15130009. Pablo Sánchez provided excellent research assistance.

†Department of Economics, Universidad de Chile.

‡Department of Economics, Universidad de Chile.

§School of Business and Economics, Universidad de los Andes.

¶Ministry of Education.

# 1 Introduction

Motivating teachers to improve the quality of instruction is a central issue in education policy. Teacher quality is highly heterogeneous, and schools face a non-trivial likelihood of hiring low-performing teachers (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014). A potential reason for this phenomenon is the fact that teacher pay, at least in the public sector, is by and large independent of performance. With the goal of improving teaching quality and motivated by insights from agency theory (Holmstrom and Milgrom, 1987, 1991), many governments have implemented pay-for-performance schemes, where teacher pay is a function of student performance in standardized test scores (Jackson et al., 2014).[1] Instead, in this paper, we assess if the same goal can be achieved by attaching teacher pay to measures of teaching skills.

This paper studies the effects of a large-scale teacher performance pay program on student academic performance. In 2016, the government of Chile implemented a teacher evaluation system associated with substantial financial incentives. We use administrative data on over 500,000 student-teacher matches, containing information on student standardized test scores on math and language skills and measures of teaching quality, and exploit the staggered implementation of the reform using a difference-in-differences approach (DiD) to identify the effects of this policy on students' academic performance. Further, we estimate a structural model to predict the impacts of alternative teaching evaluation policies. Our analysis reveals that teachers do respond to high-stakes teacher evaluation systems and that students are positively affected by them; however, the extent to which such policies are welfare-improving depends on how incentives are placed.

The education reform we study introduced the System of Teacher Professional Development (STPD), which strengthened the monetary incentives within the teacher's salary schedule. Under the STPD, teachers are placed in specific performance categories corresponding with different payment levels. Placement in these categories depends on two factors: years of teaching—such that the highest-paying categories are saved for more experienced teachers—and meeting minimum performance targets, where performance is based on a series of teaching skills measures. As a result of STPD, teacher salaries grew by almost 30% on average relative to the pre-reform pay schedule.

[1] Some examples are the teachers' pay reform in the United Kingdom (Sharp et al., 2017), the Carrera Magisterial Program in Mexico (Santibañez et al., 2007), and the System of School Performance Assessment in Chile (Contreras and Rau, 2012). See Neal (2011) for more examples.

We show that the effects of the STPD on student performance are heterogeneous and highly dependent on the policy's design. Our DiD estimates—robust to different estimation strategies and empirical samples—suggest that the STPD has negligible effects on average. Furthermore, the effects are precisely estimated, with standard errors close to $0.02\sigma$. We do not find any evidence of statistically significant pre-trends. Comparing the estimates for different definitions of our control group, we do not find evidence of anticipating behavior that could invalidate our results.

Turning into the mechanisms explaining this lack of impact, we explore two channels: (i) a low association between teacher and student tests scores and (ii) a weak effect of the policy on teacher's effort. We find stronger evidence for the second channel. First, the relationship between teacher and student tests scores is, although modest, positive. This result implies that some teachers can improve their measured skills and subsequently improve student learning. Second, we uncover positive and statistically significant results for teachers who face the larger gains per effort, in line with the hypothesis that teachers are indeed responding to the new incentives, yet these are not uniformly strong for everyone. Specifically, we find positive effects for students taught by more experienced teachers, which is consistent with the fact that performance bonuses are larger for teachers with longer tenure. Furthermore, STPD had a positive effect on students of teachers who—according to their past performance on similar exams—are neither too close or to far from the performance target. This is consistent with a model in which teachers' optimal effort in acquiring teaching skills depends on the likelihood of meeting the minimum performance thresholds defined by STPD, such that inframarginal teachers have fewer incentives to exert effort (Neal, 2011). Finally, we find similar patterns of heterogeneous effects when studying alternative measures of teaching quality, supporting the idea that teachers who face the largest monetary incentives within STPD are in fact responding to the policy.

The evidence suggesting that incentives are not powered enough for the average teacher motivates our structural analysis. We estimate a structural model of teacher effort and student achievement, and present evidence showing that alternative policies that equalize monetary incentives across teachers can increase the average effect of STPD on student test scores up to $0.11\sigma$-$0.25\sigma$, depending on the policy. Furthermore, we show that these alternative policies are also welfare-enhancing when considering provision costs and willingness to pay from both teachers and students.

By studying a system combining teacher evaluation and pay-for-performance, we relate to two

strands of literature. First, we contribute to the literature on teacher evaluation by documenting the effects of a high-stakes system. Teacher evaluation has been found to be an ineffective policy in terms of improving student learning (Jackson et al., 2014; Steinberg and Donaldson, 2016). Nevertheless, a few studies suggest that this lack of effectiveness might be related to the fact that these assessments are often low stakes. Dee and Wyckoff (2010) and Briole and Maurin (2021) show that teacher performance increases when teacher evaluation is associated with financial incentives, while Kraft et al. (2020) document that the introduction of high-stakes teacher evaluation systems in the US increased the quality of newly hired teachers.[2] Adding to this literature, we estimate the effects on student performance of similar policies, yet we further analyze how the design of the incentive system can play a major role in determining its effectiveness.[3]

This paper is also related to the literature on pay-for-performance schemes. So far, the evidence on the effects of pay-for-performance policies on student performance is mixed (Lavy, 2009; Atkinson et al., 2009; Muralidharan and Sundararaman, 2011; Goodman and Turner, 2013; Fryer and J, 2013; Gilligan et al., 2018; Mbiti et al., 2019). Neal (2011) discusses when and why these programs might not deliver the desired results. First, the vast majority of programs that include pay for performance associate their bonuses directly to student test scores, and linking payments to student performance means that policymakers must be able to control for student socioeconomic characteristics and other unobserved traits that can confound the individual performance assessment. This is done through the so-called "value-added" models. These models, however, can yield biased estimates of teacher quality (given self-selection of students into schools) while producing substantial noise in the value-added measures (Staiger and Rockoff, 2010). Furthermore, under a pay-for-performance schedule based on standardized test scores, educators have incentives to coach students in test-taking skills without necessarily improving subject matter knowledge or comprehension, and to focus their attention on those students who have the largest marginal contribution to the measure of teacher performance (Holmstrom and Milgrom, 1991; Baker, 1992; Neal, 2011). If teaching skills have positive effects on student performance and teacher assessments closely follow teacher human capital, a payment system based on teacher evaluations instead might be able to circumvent

[2] Cullen et al. (2021) find a similar effect on the teacher workforce composition after the introduction of a teacher evaluation system in Houston, Texas.

[3] Similar to our results, Bleiberg et al. (2021) find that the introduction of high-stakes evaluation reforms in the U.S. did not impact student performance, yet show suggestive evidence that program's characteristics might influence this result.

many of the potential pitfalls of current pay-for-performance schemes. In line with this argument, we empirically show that teacher assessments and student performance are related, so that re-configuring the policy's parameters can significantly improve the effects on student outcomes.

## 2 Institutional background

In 2016, the STPD fundamentally changed the structure of teachers' salaries. One of its components was a teacher evaluation system that introduced sizable financial incentives tied to teacher performance and skills in the public sector. This section describes the main aspects of this reform, focusing on how the reform impacted teachers' pay and incentives to improve their teaching skills.

**Table 1:** Main components of teacher salaries, before and after the reform

| Salary component | Formula | Before STPD | After STPD |
|---|---|---|---|
| Baseline National Salary (BNS) | $BNS$ = hours × hourly rate | ✓ | ✓ |
| Experience bonus | $[3.38\% + (3.33\% \times (\lfloor(years/2)\rfloor – 1))] \times BNS$ where $\lfloor x \rfloor$ is nearest lower integer of $x$ | ✓ | ✓ |
| Professional development bonus | $degree \times hours/44$, where $degree$ depends on whether the teachers has a professional degree or more. The value $degree$ increased with STPD. | ✓ | ✓ |
| Training assignment | $(BNS \times d \times \lfloor(years/2)\rfloor)/15$ where $d$ is the maximum percentage training | ✓ | ✗ |
| Academic Excellence Assignment | Based on Portfolio and PKT, three performance categories, paying monthly bonuses of 278, 185 and 93 dollars (inflation adjusted). | ✓ | ✗ |
| Performance categories | See Online Appendix B | ✗ | ✓ |

Notes: This table shows the most important components of the teacher salaries, before and after the STPD reform. We do not include programs that consider payments for principals (and other administrative duties), extreme geographical zone bonuses, and school-level payments.

Before the reform, public-school teacher salaries were largely unresponsive to performance. To get a sense of the counterfactual policy environment pre-STPD, Table 1 summarizes the main components of a typical teacher's salary before and after the reform. Before STPD, the most important components of a typical wage profile were the Baseline National Salary and the experience

bonus, which are both independent of an individual teacher's performance. Other salary components depended on school characteristics and whether the teacher completed any post-college short courses.[4] Notably, the previous teacher evaluation system (the Academic Excellence Assignment) had smaller financial incentives than those of the STPD and was completely voluntary.

The most important feature introduced by STDP is the performance payment. Each teacher is classified into performance categories which are associated with different financial rewards. This placement is based on (i) teacher performance and (ii) tenure. To assess the first component, the policy considers two measurements: the Subject and Teaching Evaluation Instrument (STEI) and Professional Portfolio of Pedagogical Competencies (Portfolio). These teacher assessments were designed to capture a wide array of teaching skills. STEI is a written exam that measures teaching proficiency in a certain subject. It assesses two levels of knowledge: specific subject matter knowledge and methodologies to teach those topics. The Portfolio is focused on evaluating pedagogical techniques, including a series of tasks designed to assess how the teacher can plan and execute a typical class. Part of this evaluation consists of a video-recorded class scored by external evaluators.[5] Online Appendix A provides further details on STEI and the Portfolio.

Given performance in STEI and Portfolio, the teacher is placed into one of five possible performance categories depending on her tenure. Online Appendix B documents the performance thresholds determining teacher payments. These categories are Initial, Early, Advanced, Expert I, and Expert II. Given a placement in one of these categories, a teacher has to spend a minimum of four years teaching under that category. After this period, teachers are once again evaluated and can continue to the next category only after scoring a minimum threshold, specific for each category, in either Portfolio or STEI. A teacher cannot jump from the lowest to the highest category even if she has the minimum score necessary to do so.[6] The progressive nature of the system means that the marginal benefit of effort in studying for the teachers' exams increases with teaching experience.

Figure 1 simulates an individual teacher's wage progress under the baseline policy and the STPD program. The figure assumes that a hypothetical teacher scores above the thresholds defined by

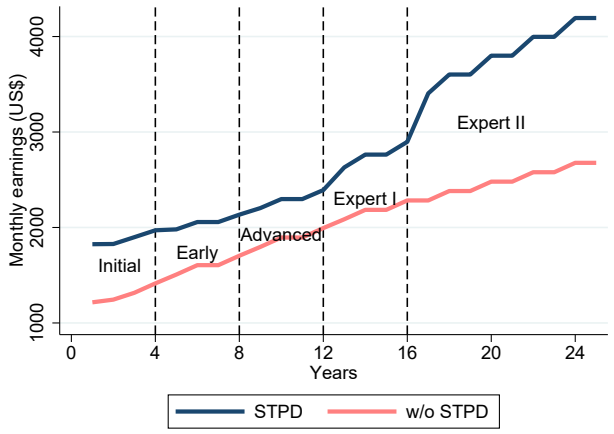
[4] In practice, the vast majority of teachers met the requirements to receive these payments.

[5] In the U.S., the National Board for Professional Teaching Standards, the Beginning Educator Support and Training Program, and the Beginning Teacher Support and Assessment Program use a similar portfolio to assess and certify teacher competencies.

[6] There are two exceptions to this rule: a teacher in the Initial can jump to the Advanced category, and a teacher in Early can go towards Expert I. To do so, the teacher must have the minimum years of experience and performance on STEI or Portfolio.

the program every time she is evaluated (every four years). When the program began in 2016, all teachers were initially placed into the performance categories based on the previous evaluation system results and on years of teaching experience. Given this initial placement, teachers then progressed onto the subsequent categories following the program's rules. Relative to the old schedule, wages under the new schedule increased by 33%. The steeper slope of the new system is notable mainly for the final two stages (Expert I and II), while for previous categories it seems that the STPD offers higher unconditional wages relative to the pre-reform era.

**Figure 1:** Simulated wage profiles with and without STPD




Notes: This figure presents two potential wage profiles. The red line shows a elementary-school teacher wages without STPD. The blue line shows the wage profile of the same type of teacher with STPD, assuming the teacher is able to score above the necessary thresholds to advance to each potential performance category: "Initial," "Early," "Advanced," "Expert I," and "Expert II."

As such, the goal of this system is to increase student performance by inducing teachers to increase their teaching skills. Nonetheless, the design of the system suggests that the monetary premiums for a unit of effort in improving STEI and/or Portfolio are not homogeneous across teachers. First, as shown in Figure 1, it is clear that more experienced teachers experienced have a larger improvement in their average wage from increasing their teaching exams, unlike the previous salary schedule. In fact, we show in Section 5 that this is consistent with our findings of positive student performance effects of the STPD associated to more experienced teachers.

Second, the fact that the system considers performance categories based on specific performance cutoffs creates weaker incentives for inframarginal teachers. To formalize this argument, suppose the production of teacher test scores follows $p = p(e, \boldsymbol{X}) + \nu$, where $e$ is effort in improving teaching skills, $\boldsymbol{X}$ is teacher fixed characteristics and $\nu$ is a random shock to teacher performance, unrelated

to $\boldsymbol{X}$, with distribution $G$. $p(0, \boldsymbol{X})$ is baseline ability, that is, the test score a teacher can produce without exerting effort. Utility is given by $U(I, e)$ where $I$ is income. To simplify the argument, suppose income has a fixed and a variable component: $I = a + \mathbf{1}\{p > c\}x^*$, meaning that a teacher gets a bonus of $x^*$ is $p > c$, where $c$ is the performance cutoff. Therefore, expected income is $a + (1 - G(c - p(e, \boldsymbol{X})))\, x^*$, and the optimal effort choice ($e^*$) is such that:

$$\frac{\partial U}{\partial I} g(c - p(e^*, \boldsymbol{X})) \frac{\partial p(e^*, \boldsymbol{X})}{\partial e} x^* = -\frac{\partial U}{\partial e},$$

where $g$ is the density of $\nu$. Optimal effort depends on how far or close is baseline ability relative to the necessary effort to meet the performance target. Suppose $p(0, \boldsymbol{X})$ is much lower than $c$. In this case, the marginal benefit is relatively low given the shape of the density $g(.)$. The same result holds for the case where baseline ability is much larger than $c$. Intuitively, inframarginal teachers have little incentives to exert effort since they know that the probability of meeting the cutoff is low (high) even if producing a large (small) effort. We provide empirical support for this theoretical prediction in Section 5 by showing that teachers in the middle of the distribution of distance to the nearest cutoff (based on previous test scores) are associated to larger positive effects of the system on student performance.

Another point to consider when anticipating the potential effects of STPD is that, even tough the policy induces teachers to increase effort, this improvement might not translate into student performance. Suppose student human capital follows $h = f(e, \boldsymbol{X})$. Even though STPD could increase $e$, if $\partial h / \partial e = 0$ then the policy will not have any impact on student performance. This problem is a version of the multi-tasking phenomenon presented in Holmstrom and Milgrom (1991), and it implies that the optimal design of a performance pay system should not include a variable component if student and teacher measures of skills are misaligned. In Section 5, we show empirically that our measures are indeed related (although this relationship is not particularly strong).

# 3 Data

To evaluate the effects of STPD on student performance, we constructed a unique data set that combines student- and teacher-level information. This section presents more details on the available

data and our estimating sample.

### 3.1 Sources and variables

Our analysis combines two administrative data sets. First, we use data on the universe of public-school teachers. This database includes all teacher-school contracts from 2005 to 2017, including information on the school where the teacher worked for all periods, the subject(s) taught, and in which classroom. Critically, knowing the teacher's classroom allows us to associate teachers with students. We restrict our analysis to 4th-grade teachers, as our observations correspond to standardized test scores for that grade level. We cannot associate teachers to a particular subject (e.g, math versus art) since teachers at that level teach multiple subjects. The data also contain the two measures of teaching quality used for categorizing teachers (STEI and Portfolio) and information on when teachers were evaluated under STPD.[7]

We pair these data with administrative records on a mandatory standardized test called Measurement System of Education Quality (Spanish acronym SIMCE). Each year, the Ministry of Education administers SIMCE to all students in particular grade levels. Our data have test scores for all students who took SIMCE between 2005-2017, administered to fourth graders.[8]. Our main outcome variables are math and language test scores for fourth-grade students. These test scores are expressed in standard deviations and are comparable across time. We only have post-reform data for fourth-graders, meaning that we only are able to estimate effects on students from that level. In addition, the Ministry surveys parents for a complete socioeconomic characterization of students' families. We use some of these variables (such as mother's education and household income) as control variables in our DiD models.

### 3.2 Sample selection

Before describing our sample restrictions, we explain a few contextual features of the reform and available data. Figure 2 illustrates the timeline of data availability and the policy's roll-out across groups of teachers. Since 2016, the STPD program has been implemented gradually across prede-

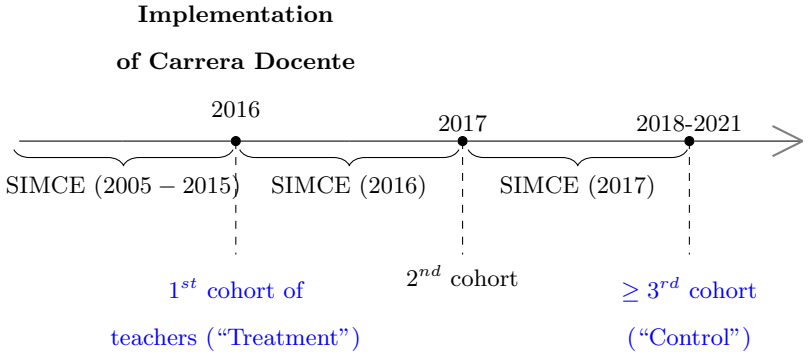
[7]For a group of teachers, our data allows us to know which year they were evaluated, for the 2016-2018 period. For a second group of teachers evaluated between 2019 and 2021, we do not observe the year in which they took the exams. We nevertheless include them in our analysis as a valid control group.

[8]SIMCE is also administered to students in one additional grade level (depending on the year, either 2nd, 8th, or 10th). We nevertheless exclude this information to have a consistent repeated cross section of only fourth graders.

termined groups of teachers with new teachers added to the program each year (recall that a teacher is again evaluated every four years). Importantly, the year before the policy was implemented none of these groups were able to determine the specific year in which they were going to be evaluated and enter the program. Only when the law was enacted, it was announced that the calendar of evaluations under STPD was going to follow the previous evaluation system schedule (which also considered a four-year window). All assessments are mandatory for public-school teachers.

We use unique identifiers to link students and teachers. Specifically, we match each teacher with test scores from fourth-grade students they taught during the years 2005-2017. We do this using the information on the school and classroom in which the teacher taught across all available years. This process gives us a repeated cross section of students and their teachers. Since SIMCE is only administered consistently to fourth graders, we restrict our data to this group of students and their associated teachers in public schools. Overall, we construct a novel pooled cross section of students with 777,832 observations (that is, student-teacher-year matches), corresponding to 771,256 students and 17,163 teachers. These numbers correspond to the total number of matches available in the original data.

**Figure 2:** Timing of implementation of STPD, data availability, and sample selection




Notes: This figure shows the years of available data, highlights the timing of initial implementation and gradual roll-out of STPD, and indicates our sample selection of treatment and control group teachers based on this timing.

We define our treatment and control groups based on the year in which the teacher was evaluated relative to the onset of STPD. We define our treatment group as students taught by teachers evaluated in 2016, the year that the STPD was implemented. Our control group are students of

teachers who were evaluated between 2018 and 2021.[9] We argue that this division captures the extent to which students and teachers were exposed to the policy. Intuitively, in 2016, teachers in the 2018 group had no (or substantially fewer) incentives to study for the teacher assessments relative to the group evaluated in 2016. To increase confidence in our identification assumption, we initially exclude students taught by teachers who were evaluated under the new system in 2017. The reason for this exclusion is that, including the 2017 group in our analysis could invalidate the identification of causal effects because of a violation of the no-anticipation condition. At the moment of implementation, teachers evaluated in 2017 knew that they were going to be evaluated that year, and they therefore could have changed their behavior (and their teaching strategy) beforehand.

Table 2 shows baseline characteristics of teachers and students in the final estimating sample. The final sample has 566,708 observations, corresponding to 562,765 students and 13,438 teachers. Panel A compares mother's and father's education of two groups of students: those who were taught by teachers evaluated in 2016 (treatment group) and by teachers evaluated in 2018-2021 (control group). These students were in fourth grade when SIMCE was administered. To compute student characteristics, we pool students for all available years. Students in our treatment and control groups come from families with relatively low levels of education, and there are no statistically significant differences between students from the treatment and control group. On the other hand, on a scale of 1 to 5, teachers score between 2.3 and 2.5 on average in their Portfolio and STEI exams. We find statistically significant differences in previous teaching experience and Portfolio. These differences (and perhaps other differences in unobserved characteristics) do not necessarily threaten our identification assumption, insofar potential outcomes in the untreated state follow a parallel evolution between groups. We formally state our identification assumption in the next Section.

[9] In some cases, more than one teacher was matched to a single classroom. In those cases, we consider an observation as treated if there was at least one teacher who was evaluated in 2016.

**Table 2:** Student and teacher characteristics across treatment groups

| | 2016 Teachers | $\geq$ 2018 Teachers | Difference |
|---|---|---|---|
| | **Panel A. Students** | | |
| Mother's schooling | 9.59 | 9.61 | -0.01 |
| | [3.23] | [3.26] | (0.04) |
| Father's schooling | 9.64 | 9.68 | -0.04 |
| | [3.56] | [3.51] | (0.04) |
| | **Panel B. Teachers** | | |
| Experience | 15.98 | 23.58 | -7.60*** |
| | [10.25] | [12.40] | (0.33) |
| Portfolio | 2.25 | 2.29 | -0.04*** |
| | [0.22] | [0.23] | (0.01) |
| STEI | 2.51 | 2.52 | -0.01 |
| | [0.44] | [0.48] | (0.02) |

Notes: This table compares observed characteristics of teachers evaluated in 2016 (treatment group) and teachers evaluated in 2018 (control group). It also compares student characteristics taught by teachers coming from these two groups. Teacher-level variables are experience (years), portfolio score, and STEI score. Student characteristics are mother's and father's schooling (years of education). We show standard deviations in square brackets and standard errors in round parenthesis. Standard errors are clustered at the school level for differences in teacher characteristics and at the teacher level for differences in student characteristics. *,**, and *** denote a statistically significant coefficient at the 10, 5 and 1% level.

# 4 Identification and estimation of causal effects

Our empirical strategy exploits the gradual implementation of the reform for different groups of teachers. This section formalizes the identification argument and presents the main empirical model to estimate causal effects.

We introduce the following notation to state our identification assumption. This notation captures the fact that we leverage from a pooled cross section of students, yet treatment occurs at the teacher level. Let $Y^z_{i(t)s}$ be the potential test score of student $i$ who was taught by teacher $t$ in period $s \in \{0,1\}$ (before or after the reform), in a counterfactual scenario where her teacher $t$ was assigned to be in the treatment group $z \in \{0,1\}$ (evaluated in 2016 or otherwise). Let $D_{i(t)}$ be an indicator for the realized treatment assignment, taking the value of 1 for students taught by teachers in the treatment group and 0 otherwise. The parallel-trends assumptions states that the average difference in student test scores produced by a teacher $t$ between those who were and were not exposed to the policy (evaluated in 2016 or otherwise), would have evolved similarly in

the absence of the reform. Formally, we assume

$$E(Y^0_{i(t)1} - Y^0_{i(t)0} \mid D_{i(t)} = 1) = E(Y^0_{i(t)1} - Y^0_{i(t)0} \mid D_{i(t)} = 0), \tag{1}$$

where $E(.)$ is taken with respect to both teachers and students.

Given the parallel trends condition above and the fact that we have a $2\times2$ design, we can identify the effect of the reform on students in the treatment group using a standard DiD regression model. Specifically, we estimate the following linear regression model:

$$Y_{i(t)s} = \beta_0 + \beta_1 Post_s + \beta_2 D_{i(t)} + \beta_3 Post_s \times D_{i(t)} + \varepsilon_{i(t)s}, \tag{2}$$

where $Post_s$ equals 1 if $s \geq 2016$ (the year the reform was implemented) and $Y_{i(t)s}$ is the observed outcome (math and language performance), such that $Y_{i(t)s} \equiv Y^1_{i(t)1} D_{i(t)} + Y^0_{i(t)1}(1 - D_{i(t)})$. Under parallel trends, $\beta_3$ identifies $E[Y^1_{i(t)1} - Y^0_{i(t)1} \mid D_{i(t)} = 1]$, the student academic performance effect of the reform on those who were assigned to a teacher who was evaluated in 2016, or the Average Treatment on the Treated (ATT). We estimate this parameter by OLS and compute standard errors clustering at the teacher level.

## 5 Difference-in-differences estimates

As discussed, the STPD injected substantial financial incentives. Agency theory suggests that such policy should induce teachers to exert more effort in acquiring more teaching skills. Furthermore, this behavioral change should have positive effects on student performance. However, while plausible, finding positive effects on students is not guaranteed. For example, it might be the case that teaching skills measures do not capture teachers' human capital properly, or that financial rewards are not sufficiently strong to change effort choices. This section presents our estimates of the STPD on student outcomes. We present our main results exploiting the comparison of students of teachers evaluated at the onset of the reform with those who were evaluated later on. We provide evidence supporting the conclusion that the reform had small-to-null effects student performance. Finally, we show evidence of heterogeneous effects, suggesting that some teachers are indeed responding to the incentives introduced by STPD.

### 5.1 Average effects

Table 3 presents our main results. It shows estimated effects of the reform on math and language test scores based on equation (2), comparing teachers evaluated in 2016 (treatment group) and in 2018-2021 (control group). We present evidence from regressions without control variables (column (1)), with student-level characteristics (column 2), and with both student- and teacher-level characteristics (column (3)). Panels A and B present estimated effects on math and language, respectively, expressed as standard deviations of the respective distributions. Most the estimated regressions show statistically insignificant effects of the reform. All DiD specifications are consistent in showing precisely estimated effects of the reform that are fairly close to zero, with estimated standard errors close to $0.02\sigma$. Only one estimate is statistically significant at the 10% level (in language, from regressions without control variables); nonetheless, the estimated impact is relatively small ($0.036\sigma$).

**Table 3:** DiD estimates of teacher reform on students' test scores

| | (1) | (2) | (3) |
|---|---|---|---|
| **A. Math (in $\sigma$s)** | | | |
| Treat×Post | 0.040 | 0.017 | 0.013 |
| | (0.024) | (0.022) | (0.023) |
| N obs. | 463,179 | 466,999 | 432,583 |
| N teachers | 13,435 | 14,520 | 13,710 |
| **B. Language (in $\sigma$s)** | | | |
| Treat×Post | 0.036* | 0.021 | 0.019 |
| | (0.021) | (0.019) | (0.020) |
| N obs. | 461,228 | 464,714 | 430,800 |
| N teachers | 13,434 | 14,499 | 13,728 |
| Student controls | | ✓ | ✓ |
| Teacher controls | | | ✓ |

Notes: This table shows DiD estimates of math (Panel A) and language (Panel B) effects of STPD. Column (1) shows baseline estimates, without control variables, Column (2) adds student-level characteristics, and column (3) includes teacher and students observed characteristics. Standard errors clustered at the teacher level are in parenthesis. *,**, and *** denote a statistically significant coefficient at the 10, 5 and 1% level.

Our available data allow us to identify effects at most two years after the teacher was first evaluated under STDP, missing some potentially important sources of effects on students. First,

it might be the case that one evaluation period is not enough to generate the desired effects on student test scores. On the other hand, individuals with larger baseline ability might decide to join the teaching profession given the steep slope in the quality-salary relationship, consistent with evidence from Tincani (2021). As such, our results are unable to speak to long-run effects coming from dynamic behavior and teacher sorting.[10] While we do not disregard the existence of such behavior, our structural analysis in Section 7 suggests that policymakers can indeed generate short-run impacts by changing the structure of incentives within STPD.

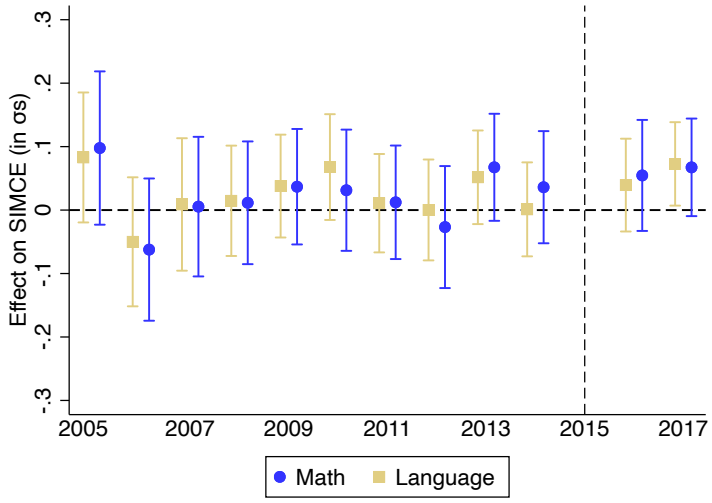
### 5.2 Threats to identification

Next, we document further evidence supporting our identification strategy and the robustness of our results. Figure 3 shows the result of a placebo test where we interact year dummies with our treatment indicator prior to the introduction of the teacher evaluation reform. This empirical exercise reveals whether there are any statistically significant pre-trends; if there are, then we would suspect that there are differences in the counterfactual evolution of tests scores for treated and non-treated students, violating the parallel trends assumption from equation (1). The figure shows that the estimated pre-trends are close to zero, precisely estimated, and statistically insignificant. This result gives us more confidence that the parallel-trends assumption is likely to hold in our data.

One potential source of bias is endogenous change in the composition of students at the teacher level. Suppose that a teacher is assigned to participate in STPD. A teacher might change schools if she thinks that her new skills have larger returns in a different setting. Likewise, suppose that the fact that teachers in a certain schools were assigned to be evaluated under STPD is common knowledge. Then, families might enroll their children to such school if they believe that would be beneficial to them. To rule out this possibility, Online Appendix C presents the same set of estimates as Table 3, with the addition of school fixed-effects. Estimates are fairly similar when we include school fixed effects, with all of the estimated ATTs remaining statistically insignificant. Furthermore, estimates are sufficiently powered to rule out moderate or even small effects, as in our main results from Table 3.

[10] Another possible effect might come from repeated teacher evaluations, independent of the associated stakes. However, de Barros (2020) shows that repeated teacher evaluations do not have effects on student learning outcomes.

**Figure 3:** Effects of teacher reform on math and language: placebo tests




Notes: The figure shows the estimation of an event-study design of the teacher reform on math and language test scores. The figure shows estimated $\gamma_j$ coefficients from the following dynamic specification:

$$Y_{i(t)s} = \beta_0 + \beta_1 D_{i(t)} + \alpha_t + \sum_{j \neq 2015} \gamma_j \mathbf{1}\{t = j\} \times D_{i(t)} + \varepsilon_{i(t)s},$$

where $j$ is year since the reform implementation and $D_{i(t)}$ the treatment group dummy. We omit teachers who were evaluated in 2017. We show 95% confidence intervals based on clustered standard errors at the teacher level.

Another element that can bias our results is a possible violation of the "no anticipation" condition. As explained in Section 2, in the year when the program evaluated a first cohort of teachers, all teachers knew the calendar of evaluations ahead. Then, it is possible that teachers who were not in the initial cohort started to study to STEI and Portfolio at the program's onset. This type of behavior could explain the null effects from our main specification (Table 3) if the students taught by teachers from the 2018 cohort were benefited from better classes thanks to this anticipating behavior.

To assess whether anticipation behavior is a problem for our setting, let us consider what would happen if we subsequently modify our control group sample under a violation of no-anticipation. Suppose that teachers are actually improving their skills only that our estimates are unable to identify it. Let us take the case of estimating the effect of the program only on 2016 student test scores and consider estimating a DiD model with three types of control groups: control group A includes all teachers (2017-2021 groups), group B includes 2018-2021 teachers (that is, our main estimates from Table 3), and group C includes 2019-2021 teachers. Note that, in the year 2016, teachers from the 2017, 2018, and 2019-2021 groups have two, three, and at least four years to prepare for the tests. If teachers need a year to prepare for the test, then we should observe a null

effect when estimating a DiD model with control group A. Nonetheless, if we use control group B instead we would obtain correct estimates and observe a positive effect. Now, suppose teachers need two years to prepare for the teacher's exams. Then, a DiD estimate would get null effects when using either control groups A or B. Nonetheless, we should get correct and positive effects of the program if we use group C as a control group. In practice, we can have a mix of teachers who take either none, one, or two years to prepare for the tests. In such case, when comparing the three types of estimates proposed here, we should observe an increasing trend in the estimated coefficients. Thus, our approach for checking whether there is an anticipation behavior is to estimate the DiD models with the three types of control groups.

Following the above argument, we first we include the 2017 teacher cohort and recover the ATT under different assumptions. We start by estimating a two-way fixed effects (TWFE) model. Let $D_{i(t)s}$ equal 1 if a student is assigned to a teacher $t$ that is in the STPD program in period $s$ and 0 otherwise. In this case, we have a staggered design: once a teacher enters STPD she never leaves. We estimate

$$Y_{i(t)s} = \beta_0 + \beta_1 D_{i(t)s} + \lambda_t + \lambda_s + u_{i(t)s},$$

where $\lambda_t$ and $\lambda_s$ are teacher and year fixed effects. Table 4 (panel A) presents the estimated TWFE coefficients on math and language. The estimates are close to zero, statistically insignificant, and have small standard errors ($0.02\sigma$).

One important caveat of our TWFE estimates is that the associated identified parameter might not coincide with the ATT. In particular, the presence dynamic heterogeneity in the effects of the program can break the identification of the average effects on the treated groups of teachers evaluated in 2016 and 2017. Under treatment-effect heterogeneity, the TWFE coefficient is a weighted average of the specific-ATT groups (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021), with weights such that the estimated coefficient of interest might not be equal to the overall ATT. Furthermore, the TWFE regression compares observed outcomes of all groups at a certain point in time, implying a comparison between just-treated and previously treated cohorts. This comparison results in negative weights, which could flip the sign of the TWFE with respect to the overall ATT. To assess if whether the estimated TWFE estimate can be of different sign, we

follow de Chaisemartin and D'Haultfœuille (2020) in two robustness checks. For math and language regressions, we find that 4% and 15% of the weights are negative, summing up to -.005 in both cases. Therefore, we conclude that the estimated TWFE coefficient is unlikely to be of different sign to that of the target parameter.

**Table 4:** DiD estimates of teacher reform on students' test scores

| | Estimate | Standard error |
|---|---|---|
| **A. Two-way fixed effects** | | |
| Math | 0.003 | 0.020 |
| Language | 0.007 | 0.019 |
| **B. Callaway and Sant'Anna (2020)** | | |
| Math | -0.006 | 0.035 |
| Language | -0.026 | 0.031 |
| **C. de Chaisemartin and D'Haultfœuille (2020)** | | |
| Math | 0.054 | 0.068 |
| Language | 0.040 | 0.059 |

Notes: This table shows estimates based on three different estimators for the effects of STPD on math and language. Panel A shows estimates of a two-way fixed effects model. Panel B shows the ATT computed using Callaway and Sant'Anna (2020). Panel C is the ATT computed following de Chaisemartin and D'Haultfœuille (2020). Standard errors are clustered at the teacher level.

Even if the TWFE regression is unlikely to yield an estimated effect of a different sign, it is still possible to obtain a non-representative coefficient if treatment effects are heterogeneous. Table 4, panels B and C, shows estimated average treatment effects following the methods by de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2020). Both methods are able to directly estimate the overall ATT, albeit under stronger parallel trends assumptions. de Chaisemartin and D'Haultfœuille (2020) estimates the overall ATT using weights for each possible group-specific DID that are proportional to the population size of such group. Callaway and Sant'Anna (2020) develop a doubly robust estimator to recover the ATT. On average, results are consistent with Table 3 in showing statistically insignificant effects. Moreover, effects are close to zero and remain precisely estimated.

Continuing with the no-anticipation analysis, we estimate the DiD model with control group C (the model with control group B corresponds to our main estimates from previous section). In this case, we estimate a simple $2 \times 2$ DiD design with 2016 teachers as treatment group and 2019-2021

teachers as control group. We report the results in Table 5. The estimates suggest that excluding the 2018 cohort does not significantly change the results as these estimates are very similar to those in Table 3.

**Table 5:** DiD estimates of teacher reform on students' test scores with 2019-2021 teachers as control Group

| | (1) | (2) | (3) |
|---|---|---|---|
| **A. Math (in $\sigma$s)** | | | |
| Treat×Post | 0.044* | 0.033 | 0.021 |
| | (0.025) | (0.024) | (0.024) |
| N obs. | 387,864 | 353,353 | 348,121 |
| N teachers | 11,147 | 11,126 | 10,992 |
| **B. Language (in $\sigma$s)** | | | |
| Treat×Post | 0.039* | 0.031 | 0.024 |
| | (0.022) | (0.021) | (0.021) |
| N obs. | 386,164 | 351,800 | 346,597 |
| N teachers | 11,148 | 11,126 | 10,992 |
| Students controls | | ✓ | ✓ |
| Teachers controls | | | ✓ |

Notes: This table shows DiD estimates of math (Panel A) and language (Panel B) effects of STPD, assuming as control group students of teachers evaluated in the 2019-2021 period. Column (1) shows baseline estimates, without control variables, Column (2) adds student-level characteristics, and column (3) includes teacher and students observed characteristics. Standard errors clustered at the teacher level are in parenthesis. *,**, and *** denote a statistically significant coefficient at the 10, 5 and 1% level.

As argued, evidence of a no-anticipation behavior should emerge by looking at estimates of our DiD models with control groups A, B, and C. Since we do not find evident trends in these estimates when we go from the models with groups A to C, we conclude that the scope for no-anticipation behavior is marginal in our setting. Note that we are not arguing that teachers do not prepare for the test. Since our panel is yearly, teachers can prepare their exams in such time frame and we would not have a violation of a no-anticipation behavior in our preferred estimate (group B). In the other extreme, we would violate the no-anticipation assumption if teachers take three years to prepare for the exam. We find this scenario unlikely, considering that completing a teaching degree takes five to six years for most undergraduate programs in Chile.

Finally, we consider a more direct evidence showing that our no-anticipation behavior is plausi-

ble. Online Appendix D provides a list of courses registered to prepare for Portfolio.[11] The longest program lasts 80 hours, much less than the prep time required for a violation of our identification assumption. All things considered, we continue with our empirical analysis using the main estimates from the previous section (control group B) to analyze the mechanisms explaining our main results.

### 5.3 Mechanisms

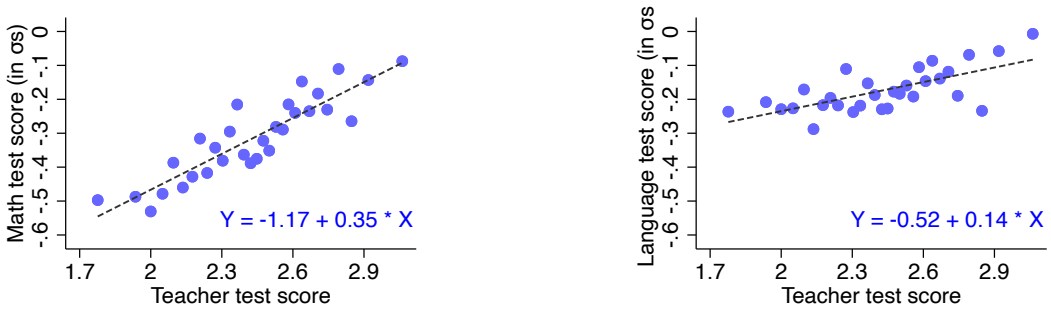
Last section showed null average effects of the policy. This section explores the mechanisms leading to such result. To this end, consider a model where the policy has effects on student outcomes indirectly through affecting teacher skills. Then, the diminished impact of the reform on student test scores can be explained by two phenomena: (i) a null causal effect of the measures of teacher skills on student outcomes and/or (ii) a null effect of the reform on teacher behavior. Either (i) or (ii) is sufficient to have null effects of the policy. For example, we could have that teachers are reacting to the policy, only that they are modifying their teaching skills in a way that is not benefiting students. In this case, the measures used in this program are not aligned with student outcomes and thus any policy configuration will not generate the desired effects on students. A second scenario is that these measures are in fact aligned to student outcomes, yet incentives are not strong enough for the average teacher. This section presents evidence supporting the second hypothesis.

Figure 4 presents correlational evidence on the relationship between teacher and student test scores. For this figure, we take all available data on teacher test scores taken before 2016. For each teacher, we average their associated SIMCE scores and estimate a linear regression on average SIMCE on past test scores. We take the average of Portfolio and STEI; if either one of these tests was not available, we consider the one that we observe for the purposes of our estimation. We do so for Math (panel a) and Language (b). The figures show a positive relationship between teacher and student test scores. The estimated regressions for math and language indicate that every additional point in teacher tests scores is associated to an increase in 0.35 and 0.14 standard deviations for math and language. This result means that a one-standard deviation increase in teacher test scores

[11] This list shows all courses are entitled to receive a subsidy under the program "Franquicia Tributaria," an on-the-job subsidy training program. Therefore, the list is not complete as there are courses which are not subsidized.

rises student test scores in 0.1 (for math) and 0.04 (for language) standard deviations, which can be considered somewhat moderate effects. While these estimates can hardly be interpreted as causal effects, we take it as evidence that teacher and student test scores are in fact related, although the relationship is not strong.

**Figure 4:** Correlation between teacher and student test scores




**(a)** Student test score: math

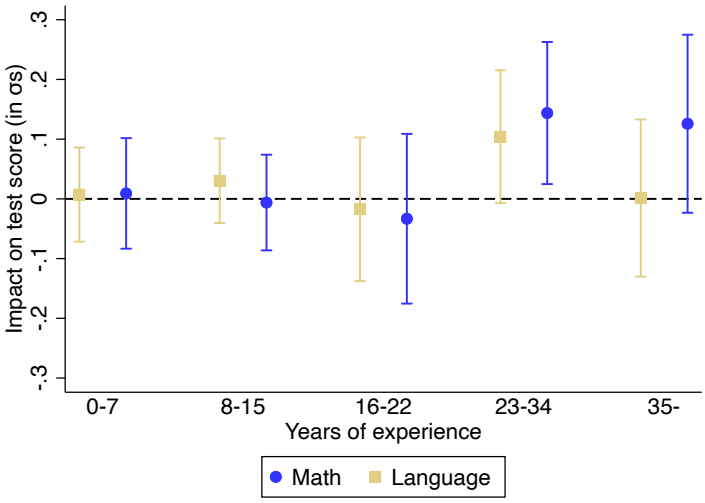
**(b)** Student test score: language

Notes: This figure shows estimated linear regressions of student and teacher test scores. Each point in the figures represents average student test score within a particular range of teacher test score. The discontinuous line is the predicted regression line of the OLS estimates.

As argued, the complexity in the wage schedule introduced by STPD implies that teachers with different characteristics might react differently to the policy. We explore heterogeneous effects in dimensions we expect to find different reactions given the design of the incentives put in place by STPD (see Section 2). First, Figure 5 presents our DiD estimates across teacher experience. In this exercise, we split the sample into five bins according to years of teaching experience: 0-7, 8-15, 16-22, 23-34, and 35 years of more. We then estimate separate regressions for each sample. The figure shows that only teachers between 23 and 34 years of experience show statistically significant and positive effects in math ($0.14\sigma$) and language ($0.10\sigma$). This result is consistent with the fact that monetary returns to improving teacher tests scores are larger for older teachers (see Figure 1). The fact that we find null results for the last category of experience might be a result of diminishing teacher productivity over time.[12]

[12]In fact, in Section 6 we find that experience has negative effects on the production of Portfolio and STEI, controlling for effort. Consistently, our model also predicts that the effects of the STPD are closer to zero for teachers near retirement (see Section 7).

**Figure 5:** Effects of STPD on math and language by years of teaching experience




Notes: This figure shows DiD coefficients from five samples based on years of experience. We show 95% confidence intervals based on teacher-level clustered standard errors.

Second, we assess if teachers with different levels of baseline abilities react differently to the policy. The fact that STPD considers minimum performance thresholds creates marginal and infra-marginal teachers, depending on the distribution of baseline teacher skills. As developed in Section 2, the optimal effort for teachers who perceive themselves as too close or far from these cutoffs in terms of their baseline skills (infra-marginal teachers) might be relatively low, since the marginal benefit of the additional effort unit is small. In contrast, teachers in the middle of the distribution of baseline skills might optimally decide to provide higher effort levels.

Figure 6 shows the effects of STPD by distance to the nearest cutoff based on previous test scores. We compute proximity based on previous STEI and Portfolio scores taken under the pre-reform evaluation system. Since for any given value of STEI and Portfolio there are two cutoffs, we compute distance for teacher $j$ as the average between the two possible distances.[13] Then, we divide the resulting sample into five quintiles of distance to the nearest progression cutoff and estimate the DiD coefficient separately for each sample.[14] However, not all teachers have prior test scores, in part due to new teacher hiring. We lose approximately 20% of observations when we condition for teachers with available past test scores.

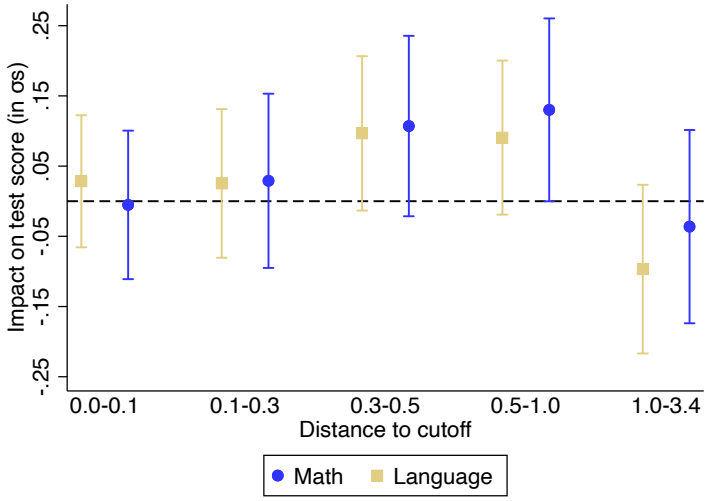
While estimates are not as precise as in our baseline estimates, we find an inverted U-shaped

[13] If a teacher can advance to a certain category by solely increasing Portfolio, then we only consider the distance to the nearest cutoff according to this measure. See Online Appendix B for more details on these performance targets.

[14] In some cases, we find that teachers, because of years of experience and high test scores, were already ranked in the highest potential category. We consider them in the first quintile of distance.

pattern in the treatment effects of the policy for both math and language. Only teachers around the middle part of the distribution of distance have positive and statistically significant effects on math and language (statistically significant at the 10% level): in the third and fourth quantiles, effects are estimated at $0.09\sigma$ and $0.13\sigma$. Moreover, the effects are consistent across math and language student assessments. Therefore, we find evidence consistent with the notion that teachers are predictably responding to incentives.

**Figure 6:** Effects of teacher reform by distance to cutoff




Notes: The figure shows difference-in-differences regressions of the teacher reform on math and language test scores by distance to the closest cutoff to advance to the next placement. We calculate teacher $t$'s distance as $d_t \equiv (s_{0t} + p_{0t})/2$, where $s_{0t}$ and $p_{0t}$ are teacher's $j$ STEI and PPCP past test scores, and $c_t^j$ is the nearest cutoff for test $j \in \{s, p\}$. We show 95% confidence intervals based on teacher-level clustered standard errors.

Online Appendix E presents evidence supporting the hypothesis that the policy improved teaching quality more for teachers facing the largest monetary incentives. The SIMCE 2016 survey included a teachers' questionnaire about different strategies used in their classes. Some examples are: "How often do you adjust your teaching plan for the month after the results of students' exams?" and "How often do you ask whether students are understanding the content during a class?" In total, we have 15 questions related to teaching methods used in class. We collapse our data at the teacher level and compare the responses in 2016 for treated and non-treated teachers according to our definition of evaluated cohorts. To reduce arbitrariness in aggregating these measures, we consider a factor model to extract one factor that we interpret as a latent teaching quality variable that underlies the correlation among the 15 questions in the survey. We average this unique factor and compare 2016 and 2018-2021 teachers across experience and distance categories. There are two

caveats worth mentioning in this analysis. First, sample size is greatly reduced as we now work with teacher-level data and not all teachers responded to the 15 questions. Second, as we only document differences in averages across groups, we do not interpret them as causal effects.[15] Notwithstanding these limitations, we find that the biggest differences in teaching quality between treatment and control groups are associated to teachers facing the largest incentives—more experienced and in the middle of the distribution of distance to the cutoff—echoing the evidence (although not exactly replicating the patterns) from Figures 5 and 6.

Online Appendix F presents tests for heterogeneous effects in two other dimensions. Figure F.1 shows estimates of equation (2) separating the sample according to initial career placement. Even though we find a statistically significant effect for teachers starting in the second level of placement, estimates are too imprecise to compare this effect against those from other categories. Figure F.2 shows effects of the policy across students' own and family characteristics. It shows that effects are statistically insignificant across student's gender and mother's education; however, we find the program had positive and statistically significant effects for math and language ($0.09\sigma$ in both cases) for students with college-graduated fathers.

Concluding, we find evidence suggesting that the null effects of the policy is explained by the differing incentives teachers face within the STPD. In spite of an apparent modest relationship between teaching skills (as measured by Portfolio and STEI) and student learning, the policy did impacted some students. Given the evidence presented in Figure 4, we cannot discard the hypothesis that teachers facing weaker incentives are actually improving their teaching skills and test scores; if they are, such improvement is not large enough to generate effects on students. In any case, our evidence support the conclusion that the way incentives were designed matters for explaining the effects of the policy on students.

[15] SIMCE's questionnaires on teaching strategies are not homogeneous across years. In particular, in 2015 questions are significantly different from those of 2016. This fact means thatwe cannot implement a DiD estimation using these particular measures of teaching skills as dependent variables.

# 6 Structural model of teacher effort, teacher evaluation, and student outcomes

Results from the previous Section indicate that the monetary incentives introduced in the STPD did not produce effects on students on average. It would be tempting then to conclude that monetary incentives do not work in the educational context. However, the fact some teachers are responding to the policy and that these are precisely the ones who face the strongest incentives suggests that there is room for improving the policy's effects on students. Next, we ask if we can increase the effectiveness of the STPD such that we can improve student learning and overall well-being. To this end, in this section, we present a structural model of teacher incentives and student human capital production. Motivated by our results from the previous section, we then take the model to the data to simulate the effects of alternative policies and assess their potential to increase student learning.

## 6.1 Model

We now present a model that formalizes the mechanisms through which STPD affects student outcomes. We draw from principal-agent models to develop a model of teacher pay-for-performance (Holmstrom and Milgrom, 1987, 1991). We complement existing theoretical frameworks of pay-for-performance based on student outcomes by considering a measure that rewards teacher performance in the teacher assessments (Baker, 1992; Neal, 2011).

The timing of teacher decisions and the main trade-offs of the model are as follows. A teacher makes a static decision, choosing effort levels in acquiring subject matter competency and teaching capabilities. Teacher efforts increase both the student's and the teacher's own human capital but potentially reduces the teacher's utility. The education authority sets up a pay-for-performance system with the goal of aligning principal (the education authority) and agents' (teachers) incentives.[16] In the system, teacher efforts is rewarded through monetary payments according to the rules of the STPD. The shape of the production function of learning, teacher's utility function, and the payment system determines optimal effort levels and learning outcomes for students.

[16] At this point, we do not analyze the problem of the optimal pay-for-performance system. Later on, we simulate the effects of implementing different, possibly sub-optimal payment schemes.

Teachers can choose effort levels from a discrete choice set that includes two types of learning efforts. They can either invest in subject knowledge—e.g., a math teacher investing in her math skills—or in pedagogical techniques—e.g, how to effectively engage students and teach subject matter. Let $e_s$ and $e_p$ denote effort towards learning subject and pedagogical skills. Both variables are binary, indicating "high" $(e_s, e_p = 1)$ or "low" $(e_s, e_p = 0)$ effort choices.

Teachers derive utility (or disutility) from different sources. Let $I$ be income and $h$ student learning. The teacher's utility function is separable in income, effort, and student learning, as follows:

$$U(I, e_s, e_p, h) = \ln(I) + \gamma_1 e_s + \gamma_2 e_p + \gamma_h \log(h^*).$$

Teacher effort choices impact teacher performance in subject and teaching skills measures. Subject and pedagogical efforts translate into changes in observed teacher performance measures. Let $M_s^*$ and $M_p^*$ be subject and pedagogical latent skills. The system of production functions that determines both skills stocks is given by:

$$\begin{aligned} M_s^* &= \alpha_0^s + \alpha_1^s e_s + \alpha_2^s experience + \alpha_3^s M_0 + u^s, \\ M_p^* &= \alpha_0^p + \alpha_1^p e_p + \alpha_2^p experience + \alpha_3^p M_0 + u^p, \end{aligned} \tag{3}$$

where *experience* is teacher's years of experience, $M_0$ is the sample mean of past test scores, and $u^j \sim N(0, \sigma^2_{M_j})$, for $j \in \{s, p\}$. Note that pedagogical effort does not affect the STEI measure, and subject effort does not affect Portfolio scores. Observed test scores are a deterministic function of latent skills. Let $M_s$ and $M_p$ be the observed performance metrics STEI and Portfolio, respectively. To match the fact that observed measures can only take values in the [1,4] scale, latent and observed measures are related given $M_j = \left[\frac{1}{1+\exp(-M_j^*)} + \frac{1}{3}\right] \times 3$ for $j \in \{s, p\}$.

Effort in subject and teaching skills translates into student learning outcomes. The production function of student skills is given by:

$$\log(h) = \beta_0 + \beta_1 e_s + \beta_2 e_p + \beta_3 experience + \nu$$

where $\nu \sim N(0, \sigma^2_\nu)$ is individual unobserved productivity at the teacher level, unrelated to effort

and experience, that affects students' observed performance. Following recent literature (Wiswall, 2013; Papay and Kraft, 2015), we allow for teacher experience to have a direct effect on student outcomes.

Teachers' salaries are determined by a complex formula that depends on a particular payment system for public-school teachers.[17] The wage schedule is thus summarized in the following equation:

$$I = (1 - D)\left[s(\boldsymbol{X}) + f(M_s, M_p, \boldsymbol{X})\right] + D\left[\widetilde{s}(\boldsymbol{X}) + \sum_{k}^{K} I_k(M_s, M_p, years)\widetilde{f}_k(\boldsymbol{X})\right], \tag{4}$$

where $D = 1$ if a teacher is under STPD and 0 otherwise. $s(D)$ and $\widetilde{s}(D)$ are baseline salaries with and without STPD. $\boldsymbol{X}$ is a set of observed teacher characteristics, such as experience, type of school, the share of vulnerable students taught by the teacher, and other variables that shift the wage schedule conditional on a level of teacher performance. For most combinations of the possible values of the components of $\boldsymbol{X}$, we have that $s(\boldsymbol{X}) < \widetilde{s}(\boldsymbol{X})$. $f(M_s, M_p, \boldsymbol{X})$ is the variable component pre-STPD. This component is a function of the teaching skills measures, capturing the fact that the old system did have an evaluation system tied to monetary rewards (although smaller than the newer system). $\sum_k^K I_k(M_s, M_p, years)\widetilde{f}_k(\boldsymbol{X})$ summarizes the performance pay component of STPD. There are $K$ possible categories, each associated with a different payment level $\widetilde{f}(\boldsymbol{X})$. $I_k(M_s, M_p, years)$ is an indicator function that equals 1 if the teacher is located in the $k$th performance category. According to the parameters of the program (see Section 2), being in a specific category depends on the the observed measures $M_s$ and $M_p$ and on teaching experience ($years$).

### 6.2 Structural estimation

In the previous subsection, we developed a model that incorporates some of the mechanisms by which STPD can have effects on student performance. This section presents our approach to taking this model to the data. We discuss (global and local) identification and our estimation strategy.

We estimate the model using the Simulated Method of Moments (Gourieroux et al., 1993). The

[17] We assume teachers have no outside options from the public sector. There are no extensive-margin labor supply responses.

procedure estimates the structural parameters by matching simulated moments with those obtained directly from the data. To implement the estimation method, we first compute the target moments from the data, $\widehat{\boldsymbol{g}}$. Then, we solve the model and obtain optimal effort choices for $M$ samples of size $N$. In these $M$ samples the structural random shocks $(\nu, e_1, e_2)$ are kept fixed. Finally, for a given vector of structural parameters $\boldsymbol{\psi}$, we compute the set of equivalent moments from the simulated choices for each of the $M$ draws. Let $\widehat{\boldsymbol{g}}^m(\boldsymbol{\psi})$ the vector of simulated moments of the $m$th draw and $\widehat{\boldsymbol{g}}(\boldsymbol{\psi}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{g}}^m(\boldsymbol{\psi})$. We obtain the the estimated parameters $\widehat{\boldsymbol{\psi}}$ as follows:

$$\widehat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \left[\, \widehat{\boldsymbol{g}} - \widehat{\boldsymbol{g}}(\boldsymbol{\psi})\right]' \boldsymbol{W} \left[\, \widehat{\boldsymbol{g}} - \widehat{\boldsymbol{g}}(\boldsymbol{\psi})\right],$$

where $\boldsymbol{W}$ is the weighting matrix. Following Blundell et al. (2016), instead of employing the efficient $\boldsymbol{W}$ matrix, we use the inverse estimated variance-covariance matrix of $\widehat{\boldsymbol{g}}$.

As is typical in these types of models, identification relies on functional form and distributional assumptions. However, the policy environment offers an additional source of identification coming from quasi-experimental variation. In the model, the STPD policy is a shock that shifts optimal effort levels for a given set of parameters determining utility and production functions. Thus, an identifying source within the model is the fact that the STPD policy does not affect preferences or the production functions, yet shifts choices by shocking the individual's salary determination.

While the policy shock provides an exogenous source of variation for global identification, all of our chosen target moments contribute to identification at the local level. Our target moments are based on data from both the 2016 and 2018-2021 groups. As the model is defined at the teacher level, we construct all moments by first averaging at such level of aggregation. Moreover, we average both SIMCE test scores for the 2016-2017 period to have a unique measure of student performance. All chosen moments hold identification power for some of the parameters of the model. First, we identify the parameters of the utility function governing the dislike of effort choice by matching simulated placements in STPD of 2018 teachers with the initial placements obtained directly from the data. The idea here is to simulate what would placement in the performance categories be for teachers who do not have financial incentives (2018 teachers); if the cost of effort is too high, then teachers would tend to exert low effort and locate in a low-performing category. On the other hand, if the teacher utility gain from higher SIMCE scores is relatively small, then student test scores

(associated with teachers with no STPD) will be low as well. Therefore, we identify the direct effect of student performance on teacher's utility by matching simulated test scores for 2018 teachers with pre-reform tests scores for 2016 teachers; here, we assume that that test scores of 2018 teachers (who do not face the STPD policies) are a good counterfactual outcome for the 2016 teacher cohort's pre-reform test scores. To locally identify the parameters of the student performance production function, we use the correlation between teacher tests scores and student performance as well as other moments of the distribution of SIMCE math test scores. Finally, to identify the parameters of the teacher production function of test scores, we use the ex-post placement of 2016 teachers after they took the 2016 Portfolio and STEI exams. Intuitively, if the productivity of effort in teacher test scores is large, then we should observe a sizable number of teachers in the highest performance categories under STPD. Online Appendix G presents the complete list of target moments and shows that the estimated model is able to predict all of these moments relatively well.

### 6.3 Estimated parameters and model validation

Table 6 presents the list of structural estimates. Panel A shows the estimated parameters of the utility function. Our estimates imply that teachers have a positive valuation of student achievement: a teacher would sacrifice 0.04 log income units for a one-standard-deviation increase in SIMCE. The two effort types have negative (direct) effects on teacher's utility; income would have to increase 0.09 and 0.07 log units to fully compensate high efforts in improving teaching and subject skills, respectively.

Panel B shows the estimated parameters of the student performance production function. Effort choices in both types of teaching skills have positive effects on students' SIMCE scores. The estimated effects are relatively large: high effort in Portfolio and STEI skills increase SIMCE by $0.30\sigma$ and $1.11\sigma$, respectively. After accounting for effort choices, the remaining variance allocated in the unobserved component is relatively small—the standard deviation of SIMCE at the teacher level is $0.04\sigma$. Teacher experience has a positive effect on student SIMCE: for every year of teaching experience, SIMCE increases by $0.007\sigma$.

Panels C and D present the estimated structural parameters of the production function of teacher performance. We find that effort has positive effects on teacher test scores. The table presents the coefficients in the latent score scale, which are difficult to read (see equation 3).

Translating these coefficients into observed scores we have that high effort in pedagogical and subject skills increase the observed Portfolio and STEI scores by 0.47 and 0.15 points in the [1,4] scale. These numbers represent a 21 and 6% rise relative to the average of past 2016 test scores (see Table 2). Therefore, increasing STEI scores is harder than improving Portfolio, even though the former has larger effects on student learning than the latter. Past test scores also have positive effects on current test scores, suggesting that teaching skills show persistence in time (conditional on effort choices and experience), where the degree of persistence is stronger for STEI. Experience has negative effects on both test scores, implying that the necessary effort investment to earn the bonus for more experienced teacher becomes larger over time.

**Table 6:** Estimated structural parameters

| Parameter | Estimate | S.E. |
|---|---|---|
| *A. Utility function* | | |
| Teaching skills (Portfolio) effort | -0.094 | 0.010 |
| Subject knowledge (STEI) effort | -0.072 | 0.009 |
| Preference for student performance | 0.041 | 0.008 |
| | | |
| *B. Production function of SIMCE* | | |
| Constant | -0.945 | 0.029 |
| Portfolio effort | 0.256 | 0.034 |
| STEI effort | 1.119 | 0.025 |
| Experience effect | 0.007 | 6.63E-04 |
| Measurement error SD | 0.037 | 0.008 |
| | | |
| *C. Production function of Portfolio* | | |
| Constant | -0.004 | 0.001 |
| Effort | 0.636 | 0.014 |
| Experience | -0.139 | 0.008 |
| Past performance | 0.042 | 0.006 |
| Variance of shock | 0.073 | 0.012 |
| | | |
| *D. Production function of STEI* | | |
| Constant | 0.000 | 4.10E-05 |
| Effort | 0.278 | 0.04 |
| Experience | -0.007 | 9.80E-04 |
| Past performance | 0.418 | 0.02 |
| Variance of shock | 0.616 | 0.02 |

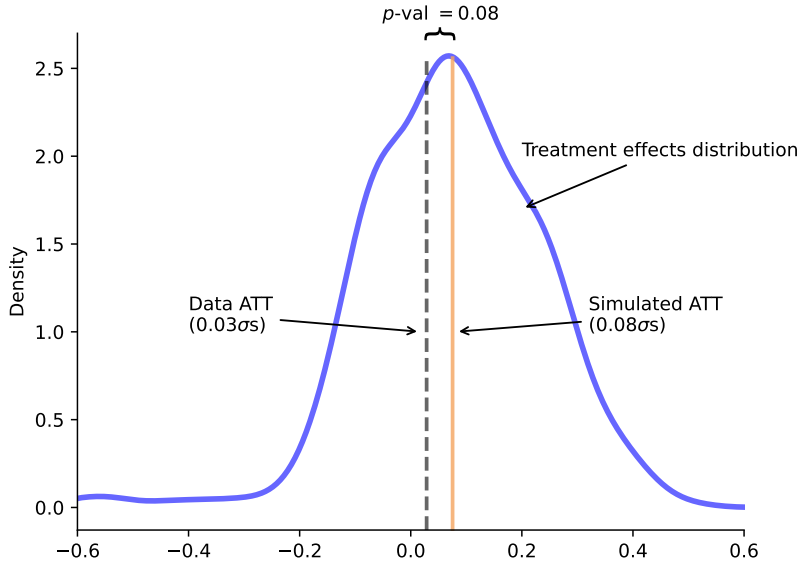
Notes: This table shows the estimated structural parameters. Standard errors are obtained via bootstrapping the entire estimation procedure (400 draws).

Together, these results suggest that teacher and student performance measures are aligned, in the sense that both are a cause of the same underlying effort choices. This alignment means that

there is there is potential to influence student performance if the policymaker provides the right incentives based on the teacher evaluation metrics.

Figure 7 provides a validation test of the model-based counterfactual estimates. It depicts the distribution of the STPD's estimated impact on students test scores as predicted by the model and compares it against the data-based (DiD) estimated ATT. To compute the model-based ATT, we focus on 2016 teachers and simulate choices and outcomes under two counterfactual scenarios: (i) salaries are determined following the STPD and (ii) salaries are determined given the pre-reform rules. We obtain ATT comparing SIMCE under these two counterfactuals. We simulate the ATT $M$ times and compute the average of those estimates for every teacher. The Figure shows that overall simulated ATT equals $0.08\sigma$, close to our preferred ATT estimation based on the DiD models ($0.03\sigma$). Furthermore, we compute a test of equality of these two estimates by bootstrapping from the original sample, and we obtain that the null hypothesis of equality cannot be rejected at the 5% level (although the associated $p$-value equals 0.08). We also find substantial heterogeneity in the model-based predicted ATT. The fact that model- and data-based estimated ATT are fairly close gives us confidence that the model is able to predict the consequences on student learning of counterfactual scenarios with reasonable predictive accuracy.

**Figure 7:** Simulated and data-based ATT of STPD on student achievement




Notes: The figure presents two estimates of the average effect of the reform on student learning: the simulated mean distribution of treatment effects estimated via the structural model, and the estimated ATT following a DiD estimation. It also shows a kernel density estimation of the simulated distribution of the individual treatment effects.

# 7 Counterfactual simulations of alternative policies

Our estimated model shows that teacher effort in subject and teaching skills affects both SIMCE and teacher performance measures. This fact gives policymakers the opportunity to reconfigure the parameters of the program with the goal of increasing the program's impact on student learning outcomes. On the other hand, our results from Section 5 suggest that making incentives more equal across teaches might boost the policy's average effect on students. This section seeks to answer the following question: can we design policies that, for a given overall cost, are more effective in terms of increasing student human capital?

## 7.1 The policies

We simulate the effects of three alternative policies and explore their effects on student performance. These policies change the marginal benefits of exerting effort, such that it make them more equals across teachers. The policies are: (i) a system without the experience conditioning, (ii) a linear pay-for-performance, and (iii) a pay-for-percentile scheme.

The first policy removes the minimum experience requirement. In this policy, if a teacher meets the minimum score to advance to a certain performance category, she is paid accordingly regardless of her total years of experience. In this way, economic incentives to improve teaching skills are now effectively homogeneous across years of teaching. Nevertheless, we keep the definition of performance categories intact.

The second counterfactual policy we study is a linear pay-for-performance scheme. We define teacher salaries following:

$$I = a + b(STEI + Portfolio)/2,$$

where $a$ and $b$ are parameters set by the education authority. We calibrate these parameters by matching two elements: (i) the minimum salary achieved by a teacher with low scores and in the beginning of her career and (ii) average wages under STDP. This policy eliminates all other salary components (see Online Appendix B), resulting in a pure pay-for-performance system for all public-school teachers. Moreover, we again disregard minimum experience requirements.

Finally, we simulate the effects of a policy where payments are based on a competition against past performance. The previous policy, while simple, presents several practical and theoretical drawbacks. First, teachers and classrooms are heterogeneous, and an optimal policy should take into account this heterogeneity. Second, from a practical point of view, the competition scheme diminishes the concern for having comparable tests across years, reducing the scope for "teaching to the test." Those issues can be tackled with a tournament scheme where payments are not directly linked to the scale of a particular test, but rather to the relative ranking a teacher can get within a contest (Neal, 2011; Barlevy and Neal, 2012).

The new counterfactual policy is based on that of Barlevy and Neal (2012). In this new system, each teacher takes Portfolio and STEI. Given these results, teachers are ranked according to the placement in the distribution of past Portfolio and STEI exams. Payments are a function of the associated percentile, conditional on a experience level. To implement the policy, we divide teachers into five experience levels: $\leq 4$, 5-10, 11-20, 21-30, $\geq 40$. Let $m = (STEI + Portfolio)/2$. For each experience level $k$, we compute $\widehat{F}_k(m)$, the cumulative distribution function of the simple average between past Portfolio and STEI, conditional on $k$. Teacher payment is then:

$$I = a + b\widehat{F}_k(m) \quad \forall k.$$

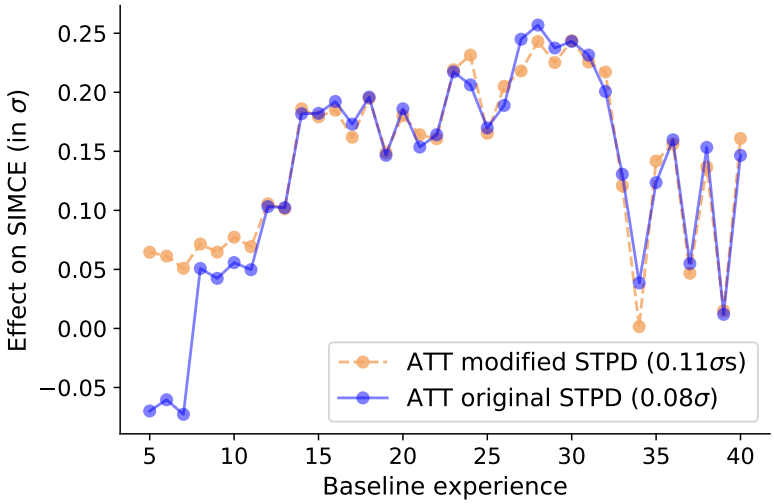
That is, teachers are "competing" against historical past performance of teachers, in contrast to Barlevy and Neal (2012) who set up contests where teachers compete against other teachers in a given cohort. We calibrate $a$ and $b$ following the same criteria as with the previous pay-for-performance policy, equalizing this new policy's minimum and average wages with those of the STPD.

### 7.2 Student outcome effects

Figure 8 presents the results from the first counterfactual experiment. It shows the simulated ATTs of the original policy with and without conditioning on teacher experience. The predicted effects of the original reform are increasing with experience levels, coinciding with the empirical estimates from Figure 5. Eliminating experience requirements raises the ATT to $0.11\sigma$, explained exclusively by the increase in the effects associated to inexperienced teachers. For these teachers, our model

predicts that the original program generates negative effects on effort and thus SIMCE, given the relatively weak incentives of the reform for this group of teachers relative to the old system.[18] While performance premiums were lower on average before 2016, they were substantially larger for young teachers as the previous payment system did not discriminate by teacher experience. Furthermore, young teachers are now receiving a larger baseline salary (see Section 2).

**Figure 8:** Simulated effects of STPD without minimum experience requirements



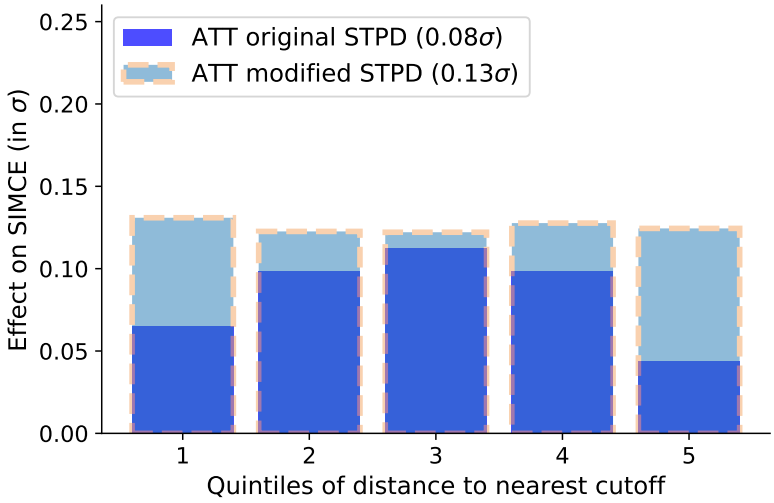

Notes: The figure plots simulated effects of two teacher pay reforms on student SIMCE. The blue scatter plot shows treatment effects of the original STPD, while the second line corresponds to the effects of a payment system where placement in performance categories do not depend on years of teaching experience. We show effects across baseline years of teaching experience.

Figure 9 shows the simulated ATT under the linear and original STPD. We compute the ATT under this new system similarly to the previous simulation. We first predict the SIMCE of 2016 teachers with and without the linear scheme, where baseline system is the one teachers faced prior 2016. We compute the ATTs across distance to the nearest cutoff following an equivalent procedure to compute distances to that of Figure 6. Relative to the estimated impact of the original STPD, we find that the linear STPD pushes the overall ATT to $0.16\sigma$. This boost comes mainly from the larger effects of teachers who were either too far or too close to meeting the cutoff; as discussed, STPD's monetary incentives for these teachers are weaker. Accordingly, the linear system equalizes marginal benefits irrespective of baseline teacher knowledge, which is the reason why the figure shows almost equal ATTs across the distribution of distance to the cutoff. Furthermore, the new

[18] In a DiD estimation (based on equation 2 and clustering standard errors at the teacher level) in the sample of teachers with eight years of experience or less, we also obtain negative effects of STPD on student math performance ($-0.09\sigma$), although the estimated coefficient is not statistically significant ($p$-value $= 0.172$).

policy has slightly larger effects for those in the middle of the distribution.

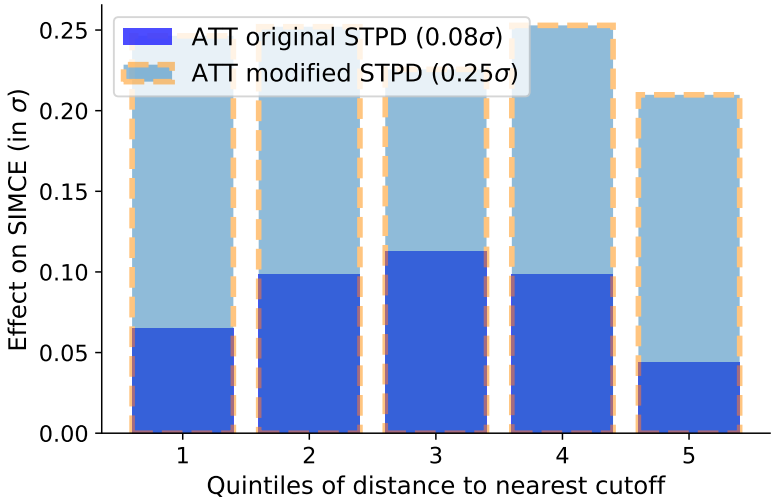
**Figure 9:** Simulated effects of STPD: original versus linear schemes across baseline teaching skills




Notes: This figure depicts simulated treatment effects of two versions of the STPD program. Blue bars show treatment effects of the original STPD. Light blue bars show effects of a linear pay-for-performance policy, where teacher salaries depend linearly on the average score of Portfolio and STEI. We show treatment effects across distance to the nearest cutoff.

Figure 10 presents the effects of the new pay-for-percentile policy on student outcomes. Of all three alternative policies considered, the pay-for-percentile policy is the one who causes the largest impact on student test scores, with an estimated effect of $0.25\sigma$. This is not to say that in all cases we should expect that a tournament of these characteristics is superior in terms of student outcomes than the linear policy. Rather, this is what we find in this particular case where policy parameters are set up to make program's costs constant. Nevertheless, we note that the pay-for-percentile scheme does condition on experience levels, which should increase the expected marginal benefits for some teachers relative to the linear pay-for-performance. On the other hand, this policy shares the virtue of the linear performance payment in that it makes incentives to put effort more homogeneous across teacher characteristics, explaining the fact that the effects on student test scores are now more similar across distance to the nearest cutoff.

**Figure 10:** Simulated effects of STPD: original versus pay-for-percentile across baseline teaching skills




Notes: This figure depicts simulated treatment effects of two versions of the STPD program. Blue bars show treatment effects of the original STPD. Light blue bars show effects of the pay-for-percentile policy, where teacher salaries depend linearly on the rank of teacher performance with respect to the historic performance on Portfolio and STEI. We show treatment effects across distance to the nearest cutoff.

### 7.3 Welfare analysis

The previous section documented relatively large improvements in the student performance effects of STPD when we change the program's parameters. However, these types of modifications to the STPD can also raise costs substantially. In this section, we study if the larger effects on student performance are valuable from a social perspective.

To make our welfare effects estimations, we compute the Marginal Value of Public Funds (MVPF) for each different policy following Finkelstein and Hendren (2020) and Hendren and Sprung-Keyser (2020). The MVPF is the ratio of the overall willingness to pay (WTP) to the net cost of the policy. The MVPF represents the dollar benefit from a marginal transfer to agents. Thus, a MVPF greater than one corresponds to a policy that generates a larger value to society than it costs (without taking opportunity costs into account). In our case, a performance pay system can drive the MVPF above the benchmark value given the incentives to improve the quality of the instruction and thus students' human capital. On the other hand, the MVPF metric allows for comparing the welfare effects across policies: if a policy A has a larger MVPF than a policy B, then policy A represents a welfare improvement relative to the counterfactual B.

To compute the MVPF we consider benefits and costs coming from choices and outcomes of both

teachers an students. Overall WTP is the sum of teachers' and students' average WTP. Students' WTP corresponds to the effect of a policy on (after-tax) life-time earnings. To have an estimate of life-time earnings, we take the average wage of students who attend public schools from Bravo et al. (2010) and assume a constant stream of earnings for 40 years (with a discount rate of 3%). Then, we follow Kline and Walters (2016) by assuming a marginal tax rate of 0.35 and that a one standard deviation increase in student performance translates into a 10% increase in life-time earnings. We obtain teachers' WTP exploiting the structure of our model: we compute the necessary annual income change that equalizes the pre-reform utility with the counterfactual policy considered.[19] Net costs correspond to provision costs minus added revenues. Provision costs equal the difference in teacher salaries between each policy to the counterfactual baseline of pre-STPD. Finally, added revenues equals the revenue collected (in present value) from the added student life-time earnings the policy generates plus teachers' added revenue coming from changes in their salaries (assuming a tax rate of 0.35).

Table 7 presents the computation of the MVPF. All of the estimates from this table are in per-capita terms. STPD has a MVPF of 0.49 in an optimistic scenario in which the policy has an effect on students of $0.08\sigma$, which is the ATT predicted by the model. This effect on student performance translates into a WTP of just \$863. On top of this, teachers' WTP is positive (\$1,030), as the policy implies larger baseline salaries. We calculate that the provision cost equals \$6,700. When we subtract added revenues from student and teachers, we obtain that net costs are \$3,956 annually. Comparing WTPs and net costs, we calculate a MVPF that is less than one. This number means that it would be more valuable to invest the same amount in a direct cash transfer to individuals.

We find that, with respect to STPD, removing the experience condition would increase welfare according to the MVPF metric. Removing the experience requirements of the policy generates larger effects on students (with a WTP of \$1,355), explaining why this policy has a larger MVFP—however, it is still less than one (0.64). Teachers, on the other hand, have a similar WTP for eliminating the experience requirement (\$979). Moreover, net costs of the policy are less than the original one, as students' added revenues increase and given the fact that provision cost are almost

[19] Changes in choices and income (given choice shifts) should not enter the WTP calculation if we consider a marginal policy (Hendren, 2016). Nonetheless, STPD and the other counterfactual policies represent substantial policy shifts. This is the reason why we use the model to compute WTP instead of just considering baseline income shifts.

constant across all experiments. All of these changes positively impact the MVPF, implying that there is an improvement in welfare when comparing this policy relative to the original STPD—however, it is still less than one (0.64).

The linear pay-for-performance scheme has larger welfare effects than the previous two policies. This policy, as we have shown in the previous section, increases student performance by $0.16\sigma$. This effect means student's WTP and added revenues are larger than both the original and modified STPD. This fact explains most of the increase in the MVPF. In this case, the MVPF is closer to the benchmark (0.95).

Finally, we compute the MVPF for the pay-for-percentile policy. In this policy, we obtain a MVPF larger than one. As in the previous case, the increase in the MVPF hinges critically on how the policy impacts student performance— this policy has the largest WTP from students ($2,573). On the other hand, this policy's MVPF (1.2) puts this type of investment at the lower bound for a set of policies directed at children reviewed in Hendren and Sprung-Keyser (2020).

**Table 7:** Marginal Value of Public Funds of counterfactual policies

| | **Original STPD** | **Policy 1 (no experience)** | **Policy 2 (linear PFP)** | **Policy 3 (percentiles)** |
|---|---|---|---|---|
| **A. Willigness to pay** | | | | |
| Students WTP (in $): $ATT \times \rho \times \$258,272 \times (1-\tau)$ | 863 | 1355 | 1802 | 2573 |
| Teachers WTP (in $) | 1030 | 979 | 1314 | 933 |
| Overall WTP | 1894 | 2334 | 3116 | 3506 |
| **B. Costs** | | | | |
| Provision cost $C$ (in $) | 6692 | 6692 | 6557 | 6600 |
| Added revenues (in $) | 2807 | 3072 | 3265 | 3696 |
| Net Cost | 3885 | 3620 | 3292 | 2905 |
| **MVPF: WTP/Net Cost** | **0.49** | **0.64** | **0.95** | **1.21** |

Notes: This table presents the estimated MVPFs for each counterfactual policy and the process for their computation. Panel A presents willingness-to-pay parameters for students and teachers. It assumes that the mapping between changes in cognitive skills and life-time earnings $(258, 272)$ is $\rho = 0.01$ (Kline and Walters, 2016). Panel B computes net costs of each policy. The provision cost $C$ equals the change in average teacher salary between each policy and the pre-reform scenario. For benefits and costs, we assume a marginal tax rate $\tau = 0.35$.

Overall, we find that welfare improvements through changes in teachers' incentives are highly

dependent on student performance effects. The break-even point for student effects that makes a policy worthwhile, given our calibrations and the estimated parameters of the model, is between $0.16 - 0.25\sigma$. Thus, we might conclude that attaching monetary incentives to teacher evaluation must generate relatively large effects on students' human capital to be welfare-enhancing. However, we view such a conclusion with caution. First, we are not considering other effects associated to students associated with improvements in performance, such as lower rates of incarceration, diminished use of welfare benefits, and so on. Moreover, we are not accounting for potential effects on non-cognitive skills, which might boost many other outcomes associated to government revenues and students' utility. Finally, we do not take into account potential effects arising in the longer-term (see section 5). As such, we consider our MVPF estimates to be lower bounds.

# 8 Conclusions

This paper studies the student effects of STPD, a policy that combines teacher evaluations with pay-for-performance. We identify the effects of the policy by exploiting the staggered implementation of the program to different groups of teachers through a DiD analysis. The bulk of our our specifications show precisely estimated null average effects of the policy. However, we document positive and economically meaningful effects for teachers who neither too close or far in terms of the likelihood of getting a wage rise. Furthermore, we show that the policy had positive effects on more experienced teachers, which coincides with the fact that the program included larger pay rises for older teachers. Building from a structural analysis, we show that a salary schedule that induces effort equally across the distribution of teacher characteristics can make the program more effective in impacting student outcomes. Furthermore, more homogeneous incentives in STPD are also welfare-improving when considering both student and teacher WTP and associated costs.

Overall, our analysis shows promising results of tying performance payments with agent's human capital development instead of agents' outputs. The evidence on pay for performance in the education literature has largely focused on programs that associate teacher payments with student performance. From a agent-principal perspective, we have shown empirically that designing performance payments based on an agent's on-the-job skills can potentially add value to an organization.

## References


Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics 25*(1), 95–135.

Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics 16*(3), 251–261.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of political Economy 100*(3), 598–614.

Barlevy, G. and D. Neal (2012). Pay for percentile. *American Economic Review 102*(5), 1805–1831.

Bleiberg, J., E. Brunner, E. Harbatkin, M. A. Kraft, and M. G. Springer (2021). The Effect of Teacher Evaluation on Achievement and Attainment : Evidence from Statewide Reforms. EdWorking Paper No. 21-496.

Blundell, R., M. Costa Dias, C. Meghir, and J. Shaw (2016). Female Labor Supply, Human Capital, and Welfare Reform. *Econometrica 84*(5), 1705–1753.

Bravo, D., S. Mukhopadhyay, and P. E. Todd (2010). Effects of school reform on education and labor market performance: Evidence from chile's universal voucher system. *Quantitative economics 1*(1), 47–95.

Briole, S. and E. Maurin (2021). There's Always Room for Improvement: The Persistent Benefits of Repeated Teacher Evaluations. Forthcoming at *Journal of Human Resources*.

Callaway, B. and P. H. C. Sant'Anna (2020). Difference-in-Differences with Multiple Time Periods. Forthcoming at *Journal of Econometrics*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

Contreras, D. and T. Rau (2012). Tournament incentives for teachers: Evidence from a scaled-up intervention in Chile. *Economic Development and Cultural Change 61*(1), 219–246.

Cullen, J. B., C. Koedel, and E. Parsons (2021). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy 16*(1), 7–41.

de Barros, A. (2020). Evaluating Teacher Evaluation – Evidence From Chile. Unpublished manuscript.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

Dee, T. S. and J. Wyckoff (2010). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management 29*(3), 451–478.

Finkelstein, A. and N. Hendren (2020). Welfare analysis meets causal inference. *Journal of Economic Perspectives 34*(4), 146–67.

Fryer, R. G. and J (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics 31*(2), 373–407.

Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal (2018). Educator Incentives and Educational Triage in Rural Primary School. NBER Working Paper No. 24911.

Goodman, S. F. and L. J. Turner (2013). The design of teacher incentive pay and educational outcomes: Evidence from the new york city bonus program. *Journal of Labor Economics 31*(2), 409–420.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics 225*(2), 254–277.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect Inference. *Journal of Applied Econometrics 8*(S1), S85–S118.

Hendren, N. (2016). The Policy Elasticity. *Tax Policy and the Economy 30*(1), 51–89.

Hendren, N. and B. Sprung-Keyser (2020). A unified welfare analysis of government policies. *The Quarterly Journal of Economics 135*(3), 1209–1318.

Holmstrom, B. and P. Milgrom (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, 303–328.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org. 7*, 24.

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics 6*(1), 801–825.

Kline, P. and C. Walters (2016). Evaluating Public Programs with Close Substitutes: The Case of Head Start. *The Quarterly Journal of Economics 131*(4), 1795–1848.

Kraft, M. A., E. J. Brunner, S. M. Dougherty, and D. J. Schwegman (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics 188*, 104212.

Lavy, V. (2009). Performance Pay and Teachers ' Effort , Productivity, and Grading Ethics. *American Economic Review 99*(5), 1979–2011.

Mbiti, I., M. Romero, and Y. Schipper (2019). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. NBER Working Paper No. 25903.

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

Neal, D. (2011). The Design of Performance Pay in Education. In E. A. Hanushek, S. J. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, Volume 4, pp. 495–550. Amsterdam: North Holland: Elsevier.

Papay, J. P. and M. A. Kraft (2015). Productivity Returns to Experience in the Teacher Labor Market. *Journal of Public Economics 130*, 105–119.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.

Santibañez, L., J. F. Martinez, A. Datar, P. J. McEwan, C. M. Setodji, and R. Basurto-Davila (2007). Breaking ground: Analysis of the assessment system and impact of mexico's teacher incentive program" carrera magisterial." technical report. RAND Corporation.

Sharp, C., M. Walker, S. Lynch, L. Puntan, D. Bernardinelli, J. Worth, E. Greaves, S. Burgess, and R. Murphy (2017). Evaluation of teachers' pay reform. Department for Education.

Staiger, D. O. and J. E. Rockoff (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives 24*(3), 97–118.

Steinberg, M. P. and M. L. Donaldson (2016). The New Educational Accountability :. *Education Finance and Policy 11*(3), 340–359.

Tincani, M. M. (2021). Teacher labor markets, school vouchers, and student cognitive achievement: Evidence from Chile. *Quantitative Economics 12*(1), 173–216.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics 100*, 61–78.

# Getting Teachers Back to School: Teacher Incentives and Student Outcomes

# Online Appendix

Patricio Araya-Córdova* Dante Contreras† Jorge Rodríguez‡ Paulina Sepúlveda§

August 19, 2022

## Contents

**A Teacher assessments** **2**

**B Performance categories and payments** **3**

**C Difference-in-differences estimates with school fixed-effects** **4**

**D Portfolio training** **5**

**E Factor analysis of teaching quality** **6**

**F Heterogeneous effects of the STPD** **9**

**G Target moments, model fit, and model validation** **11**

*Department of Economics, Universidad de Chile.
†Department of Economics, Universidad de Chile.
‡School of Business and Economics, Universidad de los Andes.
§Ministry of Education.

# A Teacher assessments

This Appendix describes the measures used to stratify teachers into the performance categories described in Appendix B. Under STPD, teachers must take two assessments: the Subject and Teaching Evaluation Instrument (STEI) and Professional Portfolio of Pedagogical Competences (Portfolio). The two tests are meant to assess knowledge dimensions specified in *Marco para la buena enseñanza* (MBE) (in English, the Good Teaching Framework). The MBE gives general guidelines for teaching best practices. It focuses on four pillars: teaching preparation, development of a good teaching environment, teacher duties, and teaching for student learning. STEI and Portfolio are designed to capture different aspects within these four dimensions. For more information on the specific content on these exams, see Rodríguez et al. (2015).

The Subject and Teaching Evaluation Instrument (STEI) measures specific academic competencies, relevant to the subject and grade level taught. It is a standardized exam consisting in two parts: multiple-choice and open questions. The first part involves 40-45 questions designed to assess the specific content each teacher should know in the particular subject she is teaching. In the second part, teachers apply the previously assessed content in a hypothetical scenario. In the case of fourth-grade teachers, individuals teach multiple subjects. Thus, STEI evaluates knowledge in "Language and Communication," "Mathematics," "History, Geography, and Social Sciences," and "Natural Sciences." In each case, the exam covers the basic content given by the official national curriculum standards.

The Portfolio evaluates teacher's performance in the classroom. To this end, teachers must submit five products to an online platform. First, they must prepare a complete plan to teach a topic relevant to the class/grades they teach. Second, they must justify an evaluation strategy. Third, they must submit a video-recorded class, with the goal of showing how well they are capable of structuring the topic in practice, understanding all topics she is teaching, and creating an effective learning environment. Fourth, teachers must submit a self-assessment analysis of their performance in the video-recorded class, taking into account factors that potentially impacted students in their learning outcomes. Finally, they must write a self-assessment essay with a critical analysis of the role they play in their day-to-day work as a teacher. These five elements are evaluated following specific rubrics that map the desired content to the 1-4 scale.

# B Performance categories and payments

This Appendix describes the determination of performance category payments. Table B.1 shows the payments associated with each performance category. For a teacher in a given initial category to advance to the following category, they must meet two conditions. First, the teacher must remain in the teaching sector for four years (until her next turn to take the teacher exams). Second, the teacher must score above STEI or Portfolio cutoffs. In the matrix from Table B.1, conditional on having the years of experience, a teacher can move up—if surpassing the Portfolio cutoff—or left—if surpassing the STEI cutoff.

**Table B.1:** Performance categories, payments, and performance thresholds

| **Portfolio cutoffs** | **STEI cutoffs** | | | |
|---|---|---|---|---|
| | 4 3.39 | 3.38 2.75 | 2.74 1.88 | 1.87 1 |
| 4<br>3.01 | Expert II<br>1,646 USD | Expert II<br>1,646 USD | Expert I<br>833 USD | Early<br>79.7 USD |
| 3<br>2.51 | Expert II<br>1,646 USD | Expert I<br>833 USD | Advanced<br>327 USD | Early<br>79.7 USD |
| 2.5<br>2.26 | Expert I<br>833 USD | Advanced<br>327 USD | Early<br>79.7 USD | Early<br>79.7 USD |
| 2.25<br>2 | Early<br>79.7 USD | Early<br>79.7 USD | Initial<br>24.1 USD | Initial<br>24.1 USD |
| 1.99<br>1 | Initial<br>24.1 USD | Initial<br>24.1 USD | Initial<br>24.1 USD | Initial<br>24.1 USD |

Notes: This table shows the payments associated with each performance category. Each cell shows a performance category as a function of Portfolio (in rows) and STEI (in columns) measures. For each measure we show the minimum and maximum scores defining each cell.

# C Difference-in-differences estimates with school fixed-effects

In this Appendix, we estimate the following model:

$$Y_{i(t,j)s} = \beta_0 + \beta_1 Post_s + \beta_2 D_{i(t,j)} + \beta_3 Post_s \times D_{i(t,j)} + \lambda_j + \varepsilon_{i(t)s}, \tag{C.1}$$

where the subscript $i(t,j)$ indicates that a student $i$ is associated with a teacher $t$ from school $j$. $\lambda_j$ is a school fixed-effect.

Table C.1 presents the estimated ATTs based on equation (C.1). This table has the same structure as Table 3, with the addition of school fixed-effects. Standard errors are clustered at the teacher level. Estimated effects are overall statistically insignificant at the 5% level. Standard errors are relatively small (around $0.02\sigma$), supporting our main conclusion of null effects.

**Table C.1:** Difference-in-differences estimates with school fixed effects

| | (1) | (2) | (3) |
|---|---|---|---|
| **A. Math (in $\sigma$s)** | | | |
| Treat×Post | 0.035* | 0.025 | 0.025 |
| | (0.019) | (0.018) | (0.019) |
| N obs. | 463,179 | 466,999 | 432,583 |
| N teachers | 13,435 | 14,520 | 13,710 |
| **B. Language (in $\sigma$s)** | | | |
| Treat×Post | 0.025 | 0.021 | 0.021 |
| | (0.017) | (0.016) | (0.017) |
| N obs. | 461,228 | 464,714 | 430,800 |
| N teachers | 13,434 | 14,499 | 13,728 |
| Students controls | | ✓ | ✓ |
| Teachers controls | | | ✓ |

Notes: This table shows difference-in-differences estimates of math (Panel A) and language (Panel B) effects of STPD. Column (1) shows baseline estimates, without control variables, column (2) adds student-level characteristics, and column (3) includes teacher and student characteristics. All estimates control for school fixed effects. Standard errors clustered at the teacher level are in parenthesis. *,**, and *** denote a statistically significant coefficient at the 10, 5 and 1% level.

# D Portfolio training

Table D.1 shows all courses registered to help preparing teachers for Portfolio. These are officially entitled to receive Franquicia Tributaria subsidy. This table shows information retrieved in https://eligemejor.sence.cl/BuscarCursoNuevo/Buscar, using the keyword "portafolio."

**Table D.1:** Providers of Portfolio courses

| Name of training course | Name of provider | Training hours |
|---|---|---|
| Estrategias Y Recursos Del Portafolio Docente | E-Syste Capacitación S.A. (E-Syste Capacitación S.A.) | 80 |
| Estrategias Para La Evaluación Y Diseño Del Portafolio Docente (on-line) | Crate Capacita Limitada (Crate Capacita ) | 32 |
| Estrategias Para La Evaluación Y Diseño Del Portafolio Docente (in person) | Crate Capacita Limitada (Crate Capacita ) | 32 |
| Diseño del Portafolio Docente | Lilian Rosario Rivas Avila Capacitación En El Trabajo E.I.R.L. | 16 |
| Aplicación De Estrategias Para La Evaluación y Diseño Del Portafolio Docente | Transversal Capacitaciones Limitada (Transversal Capacitaciones Limitada) | 16 |
| Técnicas De Elaboración Del Portafolio En La Evaluación Docente | Centro De Capacitaciones Talca Spa | 8 |
| Estrategias Para La Elaboración Del Portafolio Docente | Capacitaciones Ivama Spa (Ivamacapacita) | 8 |

Source: SENCE.

# E Factor analysis of teaching quality

This Appendix analyzes differences in alternative measures of teaching quality between teachers evaluated in 2016 relative to those evaluated in 2018-2021. We have showed that STPD have positive effects on specific groups of students, suggesting that their teachers are potentially improving their teaching skills. However, we have not shown yet direct evidence of actual changes in teaching quality. Unfortunately, we cannot use Portfolio and STEI to provide such evidence, given that these measures are not available in 2016 for all teachers, by the design of the policy. Therefore, we use alternative measures of teaching quality from the SIMCE databases.

We construct teacher-level quality measures to compare treated and non-treated teachers. For the SIMCE 2016 round, teachers were asked a series of questions regarding their teaching strategies. We list these questions in Table E.1.

To assess differences between 2016 teachers (treated) and 2018-2021 teachers (non-treated), we implement a factor analysis to aggregate these measures. In particular, we estimate the following model:

$$M_{i,k} = \alpha_{0,k} + \alpha_{1,k} f_i + \varepsilon_{i,k},$$

where $M_{i,k}$ is teacher $i$'s measured teaching skill according to measure $k$, where $k = 1, ..., 15$. The factor model states that there is a single factor $f_i$ that drives the correlation between these measures. Following standard methods in the psychometric literature (in particular, the varimax rotation method), we obtain a predicted factor at the individual level, $\widehat{f_i}$. We interpret this variable as latent teaching quality. Then, we estimate by OLS:

$$\widehat{f_i} = \beta_0 + \beta_1 D_i + u_i,$$

where $D_i$ is a dummy variable that takes the value of 1 for treated teachers and 0 otherwise. We estimate this model across sub-samples defined by experience and distance to the nearest cutoff, following the same procedure explained in Section 5.

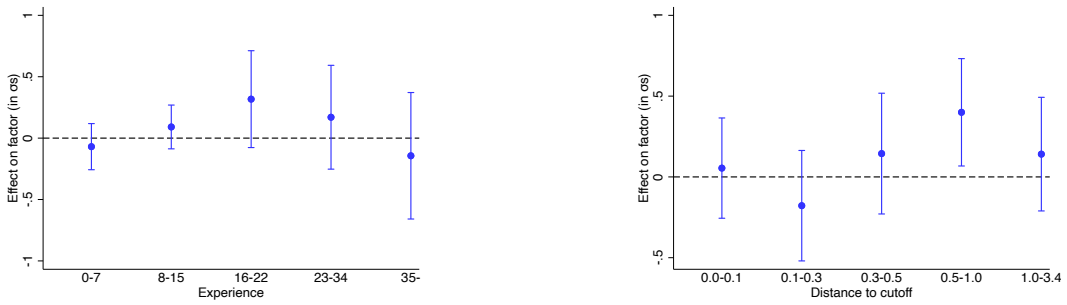
**Table E.1:** Teachers' questionnaire in SIMCE 2016

| Questions | |
|---|---|
| How often do you adjust your teaching plan for the month after the results of students' exams? | In the present year, how often have you assigned homework? |
| How often do you adjust your teaching plan for the year after the results of students' exams? | How often do you present to students the contents and objectives of a particular exam? |
| How often do you compare the results of previous student exams to assess learning progression? | How often do you detail how students will be graded before an exam? |
| How often do you adjust your teaching strategy after the results of students' exams? | How often do you present to the class common mistakes after an exam? |
| How often do you ask whether students are understanding the content during a class? | How often do you provide individual feedback after an exam? |
| How often do you present in-class practice questions to see if students have acquired the required skills? | How often do you provide ways for improving mistakes (in written) after an exam? |
| How often do you present in-class practice questions so that students can assess their own learning process? | How often do you present the solutions of an exam to the entire class? |
| How often do you present in-class practice questions so that students can assess their peers' learning process? | |

Source: SIMCE teachers' questionnaire (2016).

Figure E.1 presents the results. It shows that differences in teaching quality between treated and non-treated teachers are higher for relatively more experienced teachers and teachers in the middle of the distribution of distance to the cutoff.

**Figure E.1:** Teaching quality (alternative measures), experience, and distance to the nearest cutoff



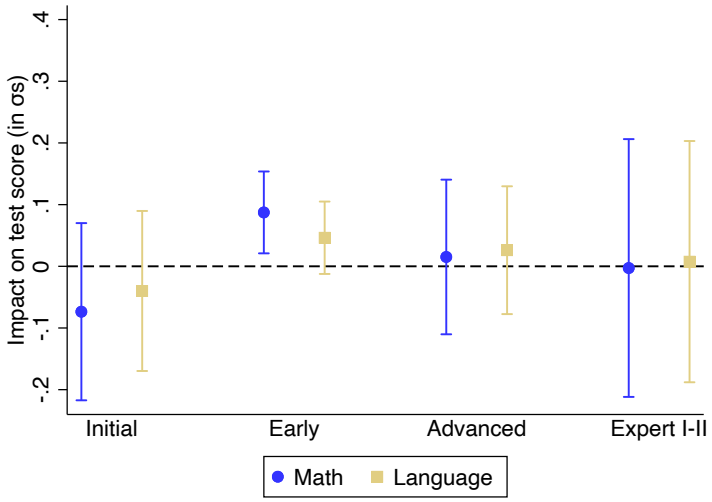

**(a)** Experience **(b)** Distance to the nearest cutoff

Notes: This figure shows averages difference between treated and non-treated teachers in the latent factor of teaching quality. Panel (a) presents differences across teacher experience and Panel (b) across distance to the nearest performance cutoff of Portfolio and STEI.

# F Heterogeneous effects of the STPD

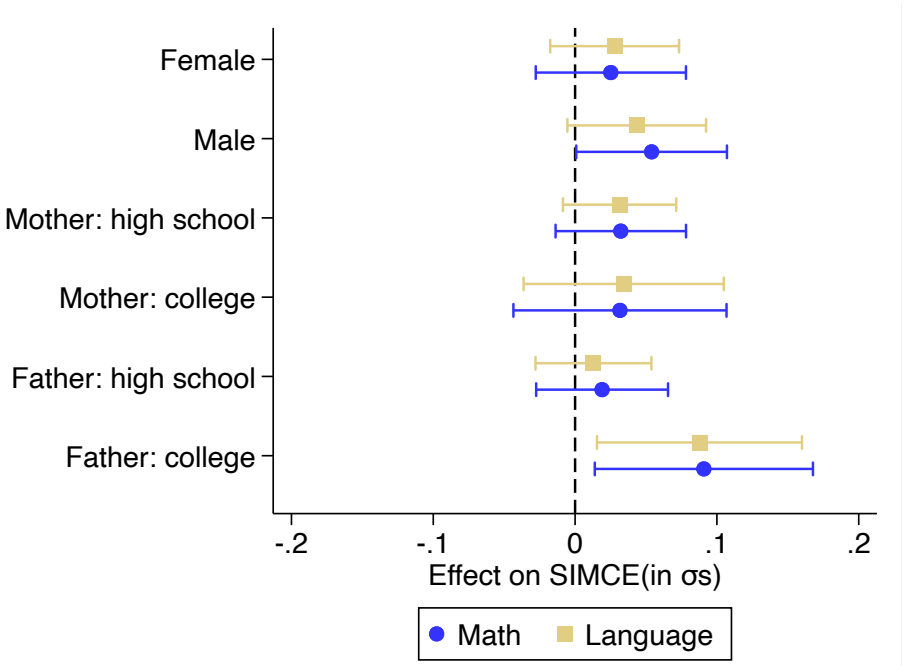
This Appendix presents further evidence on the heterogeneous effects of the STPD. Figure F.1 presents effects by initial placement. This categorization was defined when the law that implemented the STPD was enacted, and it was based on tenure and performance on exams taken prior the reform. All teachers were initially categorized following these guidelines. We estimate our regression models separating the sample according to this classification. Figure F.2 presents effects of the policy across categories of student characteristics.

**Figure F.1:** Effects of STPD by teachers' initial placement




Notes: The figure shows difference-in-differences regressions of the teacher reform on math and language test scores. It depicts estimates from four sets of regressions according to the initial placement of teachers within the program.

**Figure F.2:** Effects of STPD by student and family characteristics




Notes: The figure shows heterogeneous treatment effects of STPD on students' math and language performance. We condition on student's gender and mother's and father's education.

# G Target moments, model fit, and model validation

This Appendix presents a detailed description of the moments used in estimation and compares the moments' estimates obtained from the data and the model. Table G.1 lists all target moments and describes its computation. There are three sets of moments, Panel A lists moments exploiting the treatment group sample, Panel B describes moments using the control group, and Panel C presents the moments that use the full sample. Table G.2 compares observed (obtained directly from the data) and model-based moments. The column "Model" shows the moment computed by simulating choices and outcomes from the structural model 50 times and then taking the average of these draws. From the sample of teachers used for structural estimation, we compute the same moments directly from the data (column "data"). The final column shows the standard error of these moments computed via bootstrapping with 1,000 draws.

**Table G.1:** Definition of the moments

| Moment | Definition |
|---|---|
| **A. Treatment group (2016 teachers)** | |
| Mean Portfolio | Sample mean of Portfolio |
| Variance Portfolio | Sample variance of Portfolio |
| Mean STEI | Sample mean of STEI |
| Variance STEI | Sample variance of STEI |
| % initial | Proportion of teachers classified in the "Initial" bracket in 2016 (after receiving Portfolio and STEI results) |
| % intermediate | Proportion of teachers classified in the "Intermediate" bracket in 2016 (after receiving Portfolio and STEI results) |
| % advanced | Proportion of teachers classified in the "Advanced" bracket in 2016 (after receiving Portfolio and STEI results) |
| % expert | Proportion of teachers classified in the "Expert" bracket in 2016 (after receiving Portfolio and STEI results) |
| $corr(Port, exp)$ | Estimated correlation coefficient of teacher years of experience and Portfolio exam |
| $corr(STEI, exp)$ | Estimated correlation coefficient of teacher years of experience and STEI exam |
| % intermediate | Proportion of teachers classified in the "Intermediate" bracket when entering the program in 2016 |
| $corr(Port, past\ test\ scores)$ | Estimated correlation coefficient of Portfolio and the simple average of previous Portfolio and STEI exams |
| $corr(STEI, past\ test\ scores)$ | Estimated correlation coefficient of STEI and the simple average of previous Portfolio and STEI exams |
| **B. Control group (2018 teachers)** | |
| % advanced/expert | Proportion of teachers classified in the "Advanced" or "Expert" bracket when entering the program in 2016 |
| Average of past STEI - Portfolio | Sample mean of STEI |
| **C. Full sample** | |
| Mean SIMCE | Sample mean of the average SIMCE |
| Variance SIMCE | Sample variance of the SIMCE |
| $corr(Simce, exp)$ | Estimated correlation coefficient SIMCE and experience |

Notes: This table shows the definition of the target moments used for structural estimation.

**Table G.2:** Model fit

| Moment | Model | Data | S.E. data |
|---|---|---|---|
| A. Treatment group (2016 teachers) | | | |
| Mean Portfolio | 2.37 | 2.58 | 0.082 |
| Variance Portfolio | 0.07 | 0.06 | 0.003 |
| Mean PKT | 2.50 | 2.59 | 0.082 |
| Variance PKT | 0.25 | 0.26 | 0.010 |
| % Intermediate | 32.02 | 28.52 | 1.220 |
| % Advanced | 44.69 | 53.16 | 1.917 |
| % Expert | 19.41 | 14.19 | 0.794 |
| corr(Port,Simce) | 0.20 | 0.21 | 0.019 |
| corr(Test,Simce) | 0.21 | 0.20 | 0.020 |
| corr(exp,Port) | -0.11 | -0.06 | 0.018 |
| corr(exp,Test) | 0.01 | -0.02 | 0.018 |
| Corr(Port,p) | 0.54 | 0.58 | 0.023 |
| Corr(Test,p) | 0.12 | 0.31 | 0.020 |
| | | | |
| B. Control group (2018- teachers) | | | |
| Average of past portfolio-test | 2.43 | 2.41 | 0.076 |
| % advanced or expert | 41.81 | 31.69 | 1.054 |
| | | | |
| C. Full sample | | | |
| Mean SIMCE | -0.35 | -0.35 | 0.012 |
| Variance SIMCE | 0.26 | 0.26 | 0.009 |
| Corr(Simce,Exp) | 0.03 | 0.03 | 0.006 |
| Corr(Simce,Exp) | 0.03 | 0.03 | 0.006 |

Notes: We compare means of our target moments generated by simulating from the structural model with those estimated directly from the data.

## References

Rodríguez, B., J. Manzi, C. Peirano, R. González, and D. Bravo (2015). Reconociendo el mérito docente. Programa de Asignación de Excelencia Pedagógica, 2002-2014. Centro UC Medición - MIDE.