EdWorkingPaper No. 26-1580

# Evaluation Scores and Student Learning under Compulsory Teacher Re-Evaluation

Aaron J. Ainsworth
University of California, Irvine

Andreas de Barros
University of California, Irvine

We examine how lower-performing public school teachers respond to more frequent formative evaluations. A 2011 reform in Chile required teachers rated "basic" (the second-lowest category) to be re-evaluated after two years rather than four. Using nationwide administrative data on primary school teachers and students from 2005-2015, we show that the reform increased two-year re-evaluation rates by 58.9 percentage points. In a descriptive pre-post comparison, teachers who completed re-evaluation improved their evaluation scores by about one standard deviation. Yet, difference-in-differences estimates rule out effects on student achievement above 0.04 standard deviations, and we find no meaningful changes in student-reported teaching practices or caregiver-reported teacher caring. These results suggest that increasing the frequency of formative evaluations alone may be insufficient to remediate lower-performing teachers. They also caution against interpreting large gains in evaluation scores as sufficient evidence of improvements in teacher effectiveness.

VERSION: September 2026

Suggested citation: Ainsworth, Aaron J., and Andreas de Barros. (2026). Evaluation Scores and Student Learning under Compulsory Teacher Re-Evaluation. (EdWorkingPaper: 26-1580). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ms3w-2s74

# Evaluation Scores and Student Learning under Compulsory Teacher Re-Evaluation*

Aaron Ainsworth†

Andreas de Barros‡

July 07, 2026

We examine how lower-performing public school teachers respond to more frequent formative evaluations. A 2011 reform in Chile required teachers rated "basic" (the second-lowest category) to be re-evaluated after two years rather than four. Using nationwide administrative data on primary school teachers and students from 2005–2015, we show that the reform increased two-year re-evaluation rates by 58.9 percentage points. In a descriptive pre-post comparison, teachers who completed re-evaluation improved their evaluation scores by about one standard deviation. Yet, difference-in-differences estimates rule out effects on student achievement above 0.04 standard deviations, and we find no meaningful changes in student-reported teaching practices or caregiver-reported teacher caring. These results suggest that increasing the frequency of formative evaluations alone may be insufficient to remediate lower-performing teachers. They also caution against interpreting large gains in evaluation scores as sufficient evidence of improvements in teacher effectiveness.

**Keywords:** Chile; education; human capital; public sector labor markets; teacher evaluation.

**JEL classification:** I20, I21, I28, J45, O15.

*For support and comments, the authors would like to thank Anjali Adukia, Felipe Barrera-Osorio, Clément de Chaisemartin, Olivia Chi, José Ignacio Cuesta, Melissa Dell, David Deming, Pierre de Galbert, Alejandro Ganimian, Kathryn Gonzalez, Heather Hill, Francisco Lagos, Anne Lamb, Cristián Larroulet, María Lombardi, Eduardo Montero, Karthik Muralidharan, Lena Shi, Abhijeet Singh, Ugo Troiano, and Martin West. This paper uses proprietary data maintained by the Chilean Ministry of Education, which can be obtained by submitting a request to the Ministry (see `http://formulario.agenciaeducacion.cl/solicitud_cargar`). The authors thank Chile's *Agencia de Calidad de la Educación* and the Ministry of Education for making data available. The authors will make all code available online. Following common journal policies, the authors acknowledge the use of Artificial Intelligence (AI) for AI-assisted copy editing, yet did not use AI for generative editorial work and autonomous content creation. This paper supersedes a prior version circulated with the title "Evaluating Teacher Evaluation—Evidence from Chile". The authors have no conflict of interest to declare. The order of authors is alphabetical; co-authorship is shared equally among the authors.

†PhD Candidate, School of Education, University of California, Irvine. E-mail: `ajainswo@uci.edu`.

‡Assistant Professor, School of Education and (by courtesy) Department of Economics, University of California, Irvine. E-mail: `adb@uci.edu`.

## 1 Introduction

Public-sector labor markets face a structural challenge: even when employees are identified as underperforming, dismissals are uncommon. Employment protections, collective bargaining agreements, and administrative constraints make large-scale personnel changes unlikely. Therefore, many governments rely on alternative mechanisms to improve performance—approaches that emphasize monitoring and remediation rather than separation. This constraint has increased interest in the design of performance evaluation systems that not only identify low-performing public servants but also specify how their underperformance will be remediated.

In the public education sector, this issue is particularly salient. Teacher performance exhibits substantial variation (Jackson, 2018; Barrios-Fernández and Riudavets-Barcons, 2024), yet dismissal of low-performing teachers remains rare (Loeb, Miller and Wyckoff, 2015; Jacob, 2011). Many education systems have responded by developing evaluation systems that combine performance feedback with structured follow-up or repeat assessments. Although prior research has examined the effects of evaluations on teachers and, to a lesser extent, student learning outcomes, there is limited evidence on which features of evaluation systems influence their effectiveness. In particular, the frequency with which lower-performing teachers are re-evaluated represents an important but largely untested dimension of system design.

This paper examines a national reform that mandated an increase in the frequency of evaluations for teachers identified as lower-performing. In 2011, Chile passed a new law requiring public-school teachers ranked in the "basic" (the second-lowest) performance category of its national evaluation system to be re-evaluated after two years (instead of four). This increase in re-evaluations aimed to serve a formative purpose, with dismissal threats remaining focused on a small percentage of teachers in the (lowest) "unsatisfactory" performance category. This policy change provides a quasi-experimental setting to estimate the effects of the re-evaluation mandate for lower-performing teachers over three consecutive school years: the year of assignment to re-evaluation, the year of re-evaluation, and the year thereafter.

To measure the impact of the national re-evaluation policy, we rely on rich data sources with comprehensive coverage of Chile's education system. For the years 2005 to 2015, we use teacher-classroom links to match data on the universe of elementary teachers in the country's public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students, and results on standardized test scores for all Chilean fourth graders in mathematics and language. We further complement these data with information on teaching and teacher behaviors from student and parent surveys.

We show that the 2011 law—though imperfectly observed—sharply increased the likelihood that "basic" teachers were re-evaluated after two years (by 58.9 percentage points). This partial compliance is particularly valuable for understanding the effectiveness of the reform at scale. We then juxtapose two distinct sets of findings. First, we present descriptive statistics on teachers' evaluation scores, their associations with student- and teacher-level outcomes, and naive pre-post changes in teacher performance. Second, we employ a difference-in-differences strategy comparing "basic" teachers (subject to the new re-evaluation mandate) with more favorably rated teachers (who remained unaffected) to estimate the policy's causal effect on student- and teacher-level outcomes. To remain focused on the policy's overall effect, we report intent-to-treat (ITT) estimates.

Descriptively, teachers rated as "basic" are associated with lower student test scores, lower-quality teaching practices, and lower levels of caring. While these relationships are not necessarily causal, they lend credibility to the evaluation system's ability to identify weaker teachers in potential need of remediation. In addition, in our naive pre-post comparison, re-evaluated teachers markedly improved their evaluation scores by approximately one standard deviation (SD)—well beyond what regression to the mean can explain. Whether these gains coincide with improvements in teacher effectiveness is the question we address with our quasi-experimental analyses below.

Despite the increases to evaluation scores, our quasi-experimental findings indicate that the reform did not lead to notable improvements in student learning. The estimates are precise enough to rule out moderate positive effects (greater than 0.04 standard deviations, as per the 95-percent confidence band). If anything, most point estimates are negative, with a statistically significant negative impact of -0.055 SD in the year of a teacher's assignment to re-evaluation ($p < 0.05$). Exploratory analyses of heterogeneous effects suggest that this negative effect is driven by less-experienced teachers, female teachers, and the lowest-performing half of those rated "basic." Consistent with these null or negative effects, we also find no substantial improvements in two potentially mediating teacher-level outcomes (a teacher's instructional practices and level of caring). These findings are robust to alternative specifications and support the identifying assumptions of our difference-in-differences strategy.

Our study's first and most direct contribution lies in advancing the literature on the effects of teacher evaluation requirements on student learning. Prior literature has studied the effects of teacher evaluations on teacher-level outcomes, such as professional development activities (Koedel et al., 2019), effort (Aucejo, Romano and Taylor, 2022), job satisfaction (Koedel et al., 2017), and labor market responses (Sartain and Steinberg, 2016). In contrast, research on the effects of evaluation requirements on student learning remains limited, with only a few exceptions. Findings from these prior studies are mixed, with Taylor and

Tyler (2012), Burgess, Rawal and Taylor (2021), and Briole and Maurin (2024) documenting positive effects, yet Bleiberg et al. (2025) finding precisely estimated null effects.[1] Our study adds to this work on student-level effects by examining evaluation for remedial, formative purposes without increasing dismissal threats or awarding performance pay, providing a clean identification of this policy lever.[2]

Our study also contributes to understanding which features of evaluation systems matter. The heterogeneity in documented evaluation effects on student learning raises questions about how the specific features of evaluation systems contribute to their success or failure. Most other studies focus on the introduction of new systems or their major overhauls (as in Washington, DC, in 2009, or in Chicago, in 2012; see Dee and Wyckoff (2015) and Sartain and Steinberg (2016), respectively). In contrast, we are unaware of evidence on the causal effects of changes to a single dimension of an established teacher evaluation system, holding stakes and instruments constant. In this study, we leverage one such change—the frequency of evaluation for lower-performing teachers under Chile's national system—to estimate its causal effect on student outcomes.

Finally, our study contributes to a broader economics literature on performance measurement in settings where the measured outcome is an imperfect proxy for the objective of interest. Classic principal-agent models emphasize that evaluation based on a measured performance signal can distort effort when production is multidimensional and some dimensions are more readily measured than others (Holmström and Milgrom, 1991; Baker, 1992). In education, a related literature examines how performance is measured and the incentives those measures create (Jacob and Rothstein, 2016; Barlevy and Neal, 2012). We add to this work by examining a performance measure that is the worker's own multidimensional evaluation, the instrument a national system uses to assess its teachers, that is observed alongside an independent, test-based measure of student learning. This pairing lets us describe changes in the evaluation measure and, separately, estimate the reform's causal effect on student learning, without conflating the two.

[1] In addition, Goldhaber and Anthony (2007) find that a teacher's participation in the National Board certification process—which, like many formative teacher evaluation systems, involves preparing a teaching portfolio and external review—does not increase teacher effectiveness and reduces student learning during the certification year.

[2] Skill development and accountability are widely considered the two main channels through which teacher evaluations may improve teacher effectiveness. The reform we study operates through the formative channel: unlike the high-stakes systems studied in much of this literature, the "basic" rating carried neither performance pay nor any dismissal consequence that operated during our study period.

## 2 Context

Publicly provided primary education, governed by the Ministry of Education (*Ministerio de Educación*), serves a substantial share of students in Chile. As of 2023, public schools enrolled 36.9 percent of primary school students (The World Bank, 2025), whereas publicly subsidized private schools and other private schools educate the majority. In terms of school counts, more than half of the country's primary schools were public in 2015 (56.0 percent). Primary education enrollment rates are high, with a net enrollment rate of 94.8 percent in 2017 (The World Bank, 2025).

Chile's teacher evaluation system is a standards-based, formative assessment system that is tied to the country's national Framework of Good Teaching. In 2005, participation became mandatory for all public schools in the country. A teacher's evaluation includes four components with differing weights, as follows: a self-evaluation (10%), a third-party reference report (10%), a peer evaluator interview (20%), and a teacher performance portfolio (60%). The latter consists of a teacher's submission of a portfolio describing an eight-hour learning unit and of an announced video recording of a class. Sub-scores for each of these components are aggregated to a single, continuous performance score, which is then used to rate teachers along four performance levels: unsatisfactory, basic, competent, and outstanding.[3] The evaluation process spans one school year, and teachers receive their results and feedback at the beginning of the subsequent year (but only after they have started teaching their newly assigned classes).

In 2011, a new law (*Ley 20.501*) introduced changes to the evaluation consequences for teachers rated in the second-lowest performance category (i.e., "basic" teachers). Generally, public teachers were required to be evaluated at least once every four years.[4] Yet, under the new law, teachers rated as "basic" had to be re-evaluated after two years. For "basic" teachers, the new law did not change pay, access to incentive schemes, or the requirement to complete a professional development plan, which predated the reform; within the periods we study, the law's operative change for teachers rated "basic" (as opposed to "competent" or "outstanding") was the requirement to undergo a renewed evaluation two years later.[5]

[3]The continuous score ranges from 1 to 4, and values of 2, 2.5, and 3 are used as cut-scores, respectively. The rule-based assignment of ratings was followed closely. While a teachers final rating may be modified by a Municipal Evaluation Commission, below, we show that this phenomenon occurred in fewer than five percent of cases.

[4]The teacher evaluation system covers all teachers in municipal schools above a set workload threshold. Teachers were nominated for evaluation by the head of their respective municipal school authorities ("Municipal Education Administration Department" or "Municipal Education Corporation"). New hires were not evaluated in their first year of service. Teachers could opt out in their last three years before qualifying for retirement.

[5]Two narrower provisions of the law also singled out teachers rated "basic" or "unsatisfactory." First, employers could now require a "basic"-rated teacher to step away from classroom duties while completing

Teachers who remained “basic” after re-evaluation also faced no employment or financial consequences beyond continuing on the two-year re-evaluation cycle. Dismissal threats remained focused on teachers with an “unsatisfactory” rating, who had to be re-evaluated directly in the following year; their contracts were terminated if this rating did not improve in the re-evaluation.[6] The law came into effect in 2011, but it did not affect teachers retroactively based on their previous performance ratings. The teacher evaluation system saw further significant changes in 2016 and 2017; therefore, our study focuses on the period prior to 2016.

## 3 Data

For the years 2005 to 2015, we use teacher-classroom links to match administrative data for the universe of teachers in Chile’s public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students, results on standardized tests (in mathematics and language) for all Chilean fourth graders attending public schools, and student and parent surveys that were administered along with these exams. More precisely, we combine the following data sources.

First, to measure student learning, we use (standardized) SIMCE exam scores. The *Sistema de Medición de la Calidad de la Educación* (Education Quality Measurement System, in short: “SIMCE”) was first introduced in 1988 and represents a mandatory, full-cohort assessment. As of 2015, SIMCE had covered a wide array of subjects and levels, but most notably fourth-grade mathematics and (Spanish) language in every consecutive year since 2005. SIMCE datasets also include information on the student’s gender, school, and class, and additional student demographics (such as the mother’s highest level of education and the family’s level of income).

Second, for a subset of years, we also use survey data on two potentially mediating factors: a teacher’s teaching practices (student-reported) and level of caring (caregiver-reported).

the improvement-plan year, with the employer bearing the cost of a replacement (Art. 70). Second, school principals could propose the annual dismissal of up to five percent of a school’s teachers from among those rated “basic” or “unsatisfactory” (Arts. 7 bis and 72). Unlike the re-evaluation requirement, which applied nationwide from 2011, this dismissal authority was reserved for principals appointed through the competitive selection process that the law itself created: at the reform, no principal held it except in schools with a recent national performance-excellence award, and other schools gained covered principals only one by one, as the new process replaced sitting principals during the years we study (transitory Arts. 1, 4, and 5). Either provision, if used, would have reduced the presence of “basic”-rated teachers in classrooms; when assessing the validity of our design below, we test for and find no such pattern.

[6] In Appendix Figure A1, we show that even this threat to “unsatisfactory” teachers was weak since less than two percent of teachers received this rating. For “basic” teachers who performed “basic” (or below) in their re-evaluation, their rating could only trigger automatic dismissal in yet another, third evaluation. We do not observe any such cases in our data.

For the years 2011 through 2015, we construct an index of student-reported instructional practices. Specifically, we combine six survey questions asking students about practices their teacher engages in during class into a single index using Item Response Theory (IRT) and standardize the scores within each year in order to capture a teacher's classroom behaviors.[7] Moreover, for 2011 to 2014, we use a measure that asks caregivers to rate the extent to which their child's headteacher cares about her students.[8] As there is no head teacher identifier, we assume that in fourth grade, a teacher is considered the head teacher if she teaches both math and language.[9]

Third, we use national enrollment records for the universe of Chilean students between 2002 and 2015 to capture background characteristics. Among other variables, we observe students' absentee rate (as a percentage of school days missed), whether they repeated the grade in a given school year, and their end-of-year grade point average (GPA).

Fourth, we use yearly administrative data on the population of Chilean teachers. This data includes information on a teacher's age and gender, a teacher's years of experience in the school system, contractual details (such as the number of working hours), and information on a teacher's training (such as subject specialization and the training institution). To link this teacher-level information to students, we leverage a separate dataset with comprehensive information on the class(es) and subject(s) each teacher taught in a given year. This linking allows us to directly connect each teacher's evaluation status to student learning outcomes in the specific subjects they teach, ensuring that our estimates reflect the impact of a teacher's re-evaluation assignment on their own students' achievement.

Finally, we add data on all teacher evaluations conducted in the country between 2005 and 2015. This dataset includes detailed information on each teacher's final performance rating, the continuous performance score, and their rating on each of the four evaluation components.

[7] Students answer in four categories: "Fully agree", "agree", "disagree", "very much disagree". Students are asked about whether their teacher 1) reviews exercises, 2) reviews homework, 3) explains something repeatedly if someone asks for it, 4) continues to explain until everyone understands, 5) explains in class how tests were marked, 6) corrects the school book's exercises in class.

[8] Rated from 1 or "very unsatisfied" to 7 "very satisfied" which we standardize within year.

[9] Approximately, 80 percent of the study's sample of fourth graders have the same math and language teacher. We drop the remaining observations when analyzing mechanisms. In 2015, students were asked separately about their math and language teacher's classroom behavior. For this year, we average the student responses across subjects.

# 4 Descriptive statistics

## 4.1 Descriptive statistics on teacher evaluations

We begin by presenting descriptive statistics on whether public primary school teachers with a "basic" performance rating are associated with lower student test scores, lower-quality teaching practices, and lower levels of caring (compared to their colleagues with a higher performance rating). Appendix Figure A1 illustrates strong, monotonic positive associations between teacher evaluation scores and student learning, teaching practices, and caregiver ratings of teachers as caring. On average, students assigned to a "basic" teacher score 0.18 SD lower on the SIMCE, report teaching practices that are 0.08 SD lower, and have caregivers who rate their teachers' caring 0.09 SD lower. The distribution of evaluation scores is approximately normal, with a sizable share of teachers rated as "basic" (27.3 percent). Although purely descriptive, these results suggest Chile's evaluation ratings may serve as a useful tool for identifying those educators in need of remediation.

## 4.2 Descriptive statistics on teacher re-evaluations

Next, we present descriptive statistics on scores for re-evaluations among teachers who were assigned to re-evaluation and complied. Figure 1 confirms the above-mentioned positive associations between teacher evaluation scores and student learning also hold for re-evaluation scores. Moreover, in a naive pre-post comparison, re-evaluated teachers markedly improved their scores (by 1.01 standard deviations), a change that significantly exceeds what can be explained by regression to the mean alone.[10] Appendix Figure A2 shows the changes in scores by domain from the initial evaluation to re-evaluation. The increases in the overall evaluation scores are driven primarily by gains in the portfolio, where scores increased by 1.6 standard deviations. Scores increased in all three other areas as well, but the increases were smaller (an average of 0.49 standard deviations). Finally, students taught by teachers with stronger growth in evaluation scores showed higher test scores. We now turn from these descriptive patterns to the reform's causal, intent-to-treat effects on student learning and teacher behaviors.

[10]Under conservative assumptions, the expected regression to the mean for teachers below the re-evaluation threshold is 0.27 standard deviations (calculated using the Mills ratio method with marginal reliability of 0.8 and accounting for conditional measurement error that increases in the tails).

# 5 Empirical strategy

## 5.1 Identification

To estimate the impact of the policy, we exploit the fact that only teachers rated as "basic" were newly assigned to be re-evaluated two years later. In contrast, teachers who received a more favorable rating (i.e., "competent" or "outstanding") were only required to be re-evaluated four years later. This allows us to compare the outcomes of students taught by teachers rated basic to students taught by teachers rated above basic before and after the policy was implemented in a difference-in-differences design. Under the new policy, for a given initial evaluation year $t$, it creates a window spanning three subsequent years—$t+1$, $t+2$, and $t+3$—corresponding to the period during which "basic" teachers were assigned to re-evaluation, expected to be re-evaluated, and then followed up after re-evaluation. Prior to the policy, no teachers, including basic-rated teachers, would have been subjected to evaluation in the three years following their initial evaluation.

To assess the effects of the policy at each of these three years, we estimate the following equation separately for each period $t+1$, $t+2$, and $t+3$:

$$
\begin{aligned}
y_{ijka} = {} & \alpha_k + \tau_a + \\
& + \sum_{r \in \{\text{Unsatisf.,Basic,More fav.}\}} \gamma_r \text{Rating}_{jt} + \zeta \text{Evaluated}_{jt} \times \text{Post}_t \\
& + \sum_{r \in \{\text{Unsatisf.,Basic}\}} \beta_r \text{Rating}_{jrt} \times \text{Post}_t + \varepsilon_{ijka}
\end{aligned} \tag{1}
$$

In Equation 1, $y_{ijka}$ denotes the outcome for student $i$, taught by teacher $j$ in commune $k$ in year $a$, where $a \in \{t+1, t+2, t+3\}$ varies with the period being analyzed.[11] The term $\alpha_k$ represents commune fixed effects, and $\tau_a$ denotes year fixed effects. The term $\text{Rating}_{jrt}$ is an indicator for whether teacher $j$ received rating $r \in \{\text{Unsatisfactory, Basic, More favorable}\}$ in initial evaluation year $t$, $\text{Post}_t$ is an indicator equal to one if year $t \geq 2011$, denoting the post-policy period, and $\text{Evaluated}_{jt}$ indicates whether teacher $j$ was evaluated in year $t$. The omitted category consists of teachers who were not evaluated in year $t$.

As we estimate the policy's effects separately for each period ($t+1$, $t+2$, or $t+3$), to avoid comparisons with observations from the same evaluation year $t$ but in the other two post-evaluation periods, we drop those remaining observations from the estimation. We

[11] Communes are the smallest administrative and political subdivisions in Chile, roughly equivalent to municipalities or townships. Because public-school teachers were legally employed by municipalities, which convened and oversaw the annual teacher evaluation process for eligible teachers in their communes, we include commune fixed effects, corresponding to the level at which assignment was stratified (i.e., assignment-stratum fixed effects).

also drop all observations for students taught by teachers currently being evaluated in year $t = 0$.[12] In each of the three separate estimations, the coefficient of interest is $\beta_{\text{Basic}}$, which captures the period-specific intent-to-treat effect of assigning teachers rated as "basic" to be re-evaluated two years later under the new policy; we ignore the results for $\beta_{\text{Unsatisfactory}}$.

In estimating Equation 1, standard errors are clustered at the teacher level to account for the fact that treatment is assigned at that level (Abadie et al., 2023).

### 5.2 Compliance

First, we investigate compliance with the re-evaluation policy. Table 1 shows that, prior to the law, no primary school teachers rated as "basic" were re-evaluated within three years of their initial evaluation. In contrast, while there is non-compliance, the policy sharply increased re-evaluations two years after a teacher's "basic" rating (by 58.9 percentage points). As prescribed, these re-evaluations are almost entirely concentrated in the second year, with virtually none occurring in the remaining two years.

### 5.3 Internal validity

Table 2 describes the sample and supports the validity of our identification strategy. Panel A shows that municipalities rarely vetoed a teacher's "basic" rating if their evaluation score suggested it, and that the new law did not lead to additional systematic manipulation of this rating.

Panels B to D of Table 2 suggest the policy did not alter the extent of systematic sorting of lower-performing teachers to (or from) students. Panel B shows that, because of the policy, "basic" teachers did not become less (more) likely to teach a fourth-grade classroom whose students were required to take the national SIMCE exam. Panel C focuses on those fourth-grade students who took the exam and also confirms that the policy did not lead to a noticeable shift in the observable student characteristics taught by "basic" teachers, two years after a teacher's performance rating. Similarly, Panel D shows that the characteristics of their teachers remained largely unaltered. Global $F$ tests provide no evidence of systematic sorting ($p = 0.393$ for student characteristics; $p = 0.660$ for teacher characteristics). Appendix Table A1 reports analogous results for the other periods ($t + 1$ and $t + 3$). Two variables, student attendance and teacher contracted hours, show negligible differences for period $t + 2$, and there may be a slight imbalance in the remaining two

[12]For those teachers that are re-evaluated as a result of the policy, we do not count their new evaluation results towards treatment assignment. In other words, we only rely on initial evaluations from the typical four-year evaluation cycles.

periods. These differences are small, with the differences in student attendance of 0.69 percent representing about one fewer day of school attended. Yet, the variables closest to student performance remain balanced throughout (students' grade point average, GPA, and grade repetition), and our robustness checks (reported below) confirm that controlling for student and teacher background characteristics does not alter our results.

# 6 Results

## 6.1 Main results: Effects on student learning

Our intent-to-treat estimates indicate the policy had null or negative effects on student learning (Table 3, Panel A). Pooled across the two subjects, assigning a "basic" teacher to be re-evaluated led to -0.055 SD lower student performance in the year of a teacher's assignment ($p = 0.049$), -0.019 SD lower performance in the year of the re-evaluation (effects exceeding 0.041 SD ruled out as per the 95-percent confidence interval), and -0.033 SD lower performance in the year after the re-evaluation (effects exceeding 0.043 SD ruled out as per the 95-percent confidence interval). These findings are qualitatively similar across the two subjects, with slightly stronger (i.e., more negative) impacts in mathematics in the year after the re-evaluation.

## 6.2 Secondary results: Effects on teaching practices and caring

Consistent with the lack of positive effects on student learning, we also do not find substantial improvements in our two teacher-level outcomes (Table 3, Panel B). In the year after a teacher's assignment, we can rule out positive intent-to-treat effects larger than 0.04 SD on teachers' level of caring, and of 0.05 SD on teaching practices (as per the 95-percent confidence interval). For the remaining two years, the coefficients are close to zero or slightly positive (the largest coefficient being 0.048 SD, for the impact on teaching practices in the year of the re-evaluation; $p = 0.194$). Due to the reduced sample size for our teacher-level measures, we cannot rule out meaningful positive effects for these two years. Nonetheless, with coefficients close to zero and below 0.05 SD throughout, we cautiously conclude that, if any positive teacher-level effects exist, they were likely too modest to translate into impacts on student learning.

### 6.3 Exploration of subgroup effects

We present additional, exploratory subgroup analyses for three sets of teacher characteristics (Figure 2). We focus these analyses on the study's main outcome (student learning).[13]

Prior work suggests that the productivity of teachers with fewer years of work experience may be more malleable and, thus, teacher evaluations may be more helpful for less-experienced teachers (Chetty et al., 2011; Papay and Kraft, 2015; Bell et al., 2025). Following conventions from this literature on the returns to teacher experience, we subset the sample and repeat our analyses for a group of teachers with up to 10 years of work experience ("novice teachers") as well as for another group with more than 10 years of experience ("experienced teachers").

In addition, prior literature suggests female teachers may be more receptive to feedback from evaluations (Buurman et al., 2020). Therefore, we also repeat our analysis for subsets of female and male teachers, respectively.

Finally, as the new policy aimed to remediate the performance of lower-performing teachers, we hypothesized that, if successful, the re-evaluation mandate would be more effective for those teachers with the lowest performance. Accordingly, we repeat our analyses twice more, removing the bottom and top halves of the "basic" teachers in each iteration, based on their initial evaluation scores.

Our findings suggest the policy's negative impacts on student learning in the year of a teacher's assignment to re-evaluation are driven by less-experienced teachers, female teachers, and the lowest-performing half among those rated "basic." The point estimates, presented in Figure 2, are -0.111, -0.068, and -0.117 ($p = 0.012$, $p = 0.028$, and $p = 0.001$, respectively), with no statistically significant impacts for the remaining subgroups in that period. For novice teachers, we also find negative impacts in the remaining two periods, of -0.112 SD for $t + 2$ and -0.114 SD for $t + 3$ ($p = 0.021$ and $p = 0.062$, respectively). We do not find other, additional negative subgroup effects for these two periods.

We caution that these exploratory subgroup analyses are noisy, and we cannot rule out that there are no differences in impacts across these subgroups. However, even if suggestive, we find negative or null findings precisely for those subgroups for which we would have expected the new law to be *more* beneficial. Taken together, these exploratory findings are consistent with the earlier result that the policy did not improve student learning.

[13] For completeness, in the Appendix materials, we also present (but do not further discuss) complementary analyses for the two potentially mediating teacher-level outcomes (Appendix Figure A3); we also present the same information depicted in Figure 2 and Appendix Figure A3 as a table (Appendix Table A3).

### 6.4 Additional validity and robustness checks

Complementing our earlier checks regarding the lack of systematic endogenous sorting of assigned teachers to (from) students, we present four additional validity and robustness checks.

First, while the parallel trends assumption of our difference-in-differences strategy is untestable, we present evidence against the presence of differential pre-trends, overall (Appendix Figure A4) and for our exploratory subgroup analyses (Appendix Figures A5 and A6). As expected, in the pre-period, we do not observe a systematic trend in the test-score differences of students taught by "basic" vs. higher-rated teachers (whether one, two, or three years after a teacher's initial performance rating).

Second, we conduct a falsification test, investigating "impacts" on test scores in the year prior to a teacher's "basic" rating (Appendix Table A2, Column 1). We do not find evidence of systematic differences in test scores in the year prior to a teacher's treatment assignment.

Third, our results are robust to the inclusion of student- and teacher-level controls (Appendix Table A2, Columns 2 to 4). The resulting point estimates are similarly precise and very close to those reported in Table 3. If anything, some of the coefficients are slightly more negative, including the effect for students' language scores in the year of a teacher's assignment to re-evaluation (-0.064 SD; $p = 0.024$).

Fourth, motivated by the recent literature on two-way fixed models, we also present results from a robustness check that implements the two-stage imputation estimator developed by Gardner (2022). Our results, presented in Columns 5 to 7 of Appendix Table A2, are qualitatively similar to those obtained earlier. The point estimates for effects on the study's main outcome of interest (student learning, as measured by stacking SIMCE test scores) are within 0.01 SD ($t+1$), 0.01 SD ($t+2$), and 0.02 SD ($t+3$) of those presented earlier, in Table 3.

## 7 Conclusion

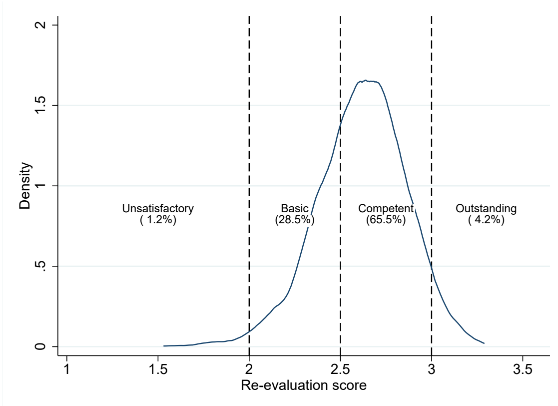
This study provided quasi-experimental evidence on the effects of a national reform that increased the frequency of evaluations for lower-performing teachers in Chile's public education system. These formative re-evaluations were expected to remediate lower-performing teachers by focusing on skill development channels rather than on changes to teacher incentives. Evaluation scores in Chile's system are positively associated with student test scores. Moreover, in a simple pre-post comparison, re-evaluated teachers markedly improved their evaluation scores upon reassessment. These gains in evaluation scores were, in turn, positively associated with student learning.

Overall, however, we found no evidence that more frequent evaluations improved student learning in the year of or following the evaluation. Our estimates are sufficiently precise to rule out positive effects on student test scores (exceeding 0.04 standard deviations, which is a small effect size for educational interventions (Kraft, 2020; Evans and Yuan, 2022)). Rather, we find assigning a teacher to re-evaluation led to a -0.05 SD negative effect on student learning in the year the assignment was announced.

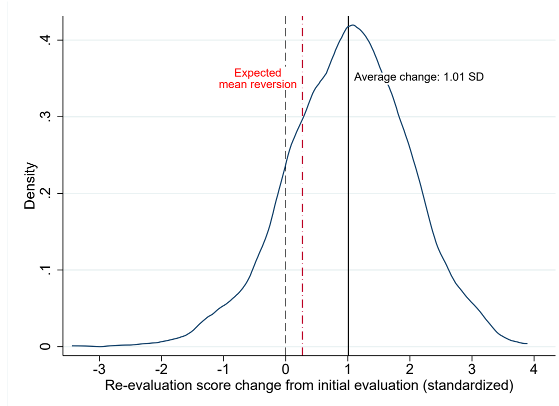
Taken together, this study provides the first causal evidence on the frequency with which lower-performing teachers are re-evaluated within a comprehensive, national system. By isolating the impact of repeat evaluations for lower-performing teachers—rather than the introduction of a new evaluation system for all teachers—this study contributes new evidence on a dimension of policy design that has received little causal attention. The results caution policymakers that solely increasing evaluation frequency may not improve the effectiveness of weaker educators and could even have unintended consequences. They also caution against interpreting improved evaluation scores after repeat assessment as sufficient evidence of improved teacher effectiveness: in this setting, large subsequent gains in evaluation scores coincided with no improvement in student learning or measured teacher behaviors. The findings offer important lessons for education systems considering teacher evaluation reforms and, more broadly, for the design of systems aimed at remediating lower-performing public servants.

While our empirical setting does not permit us to identify mediating pathways (a task beyond the scope of this Short Communication), two features of our results are consistent with potential mechanisms that future work could examine. First, the negative effects emerge immediately in the year the re-evaluation assignment is announced—before teachers undergo the evaluation process. This timing suggests the evaluation burden itself may not be the primary driver, unlike in portfolio-based certification processes, where teacher effectiveness declines during preparation (Goldhaber and Anthony, 2007). Instead, the immediate negative effect is more consistent with the re-evaluation assignment functioning as a negative signal about performance, potentially undermining teacher confidence or motivation (Koedel et al., 2017). Second, the negative effects are concentrated among subgroups expected to benefit more from formative feedback: less-experienced teachers, female teachers, and the lowest-performing "basic" teachers. Teachers in these subgroups may be particularly susceptible to motivational shocks from negative performance signals.

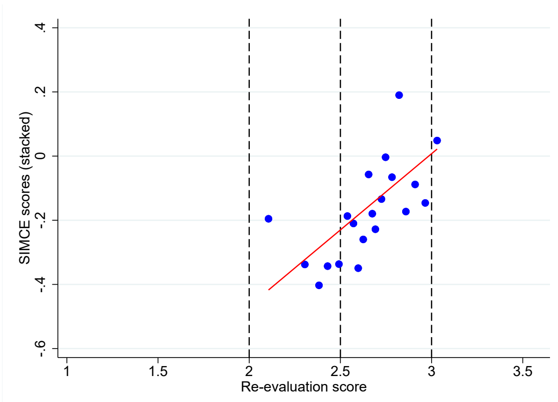
**Figure 1:** *Distribution of teacher re-evaluation scores, growth, and associations with learning outcomes*



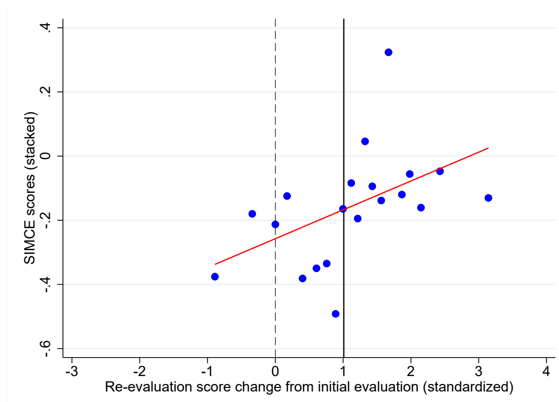

**(a)** Distribution of teacher re-evaluation scores




**(b)** Distribution of changes in evaluation scores



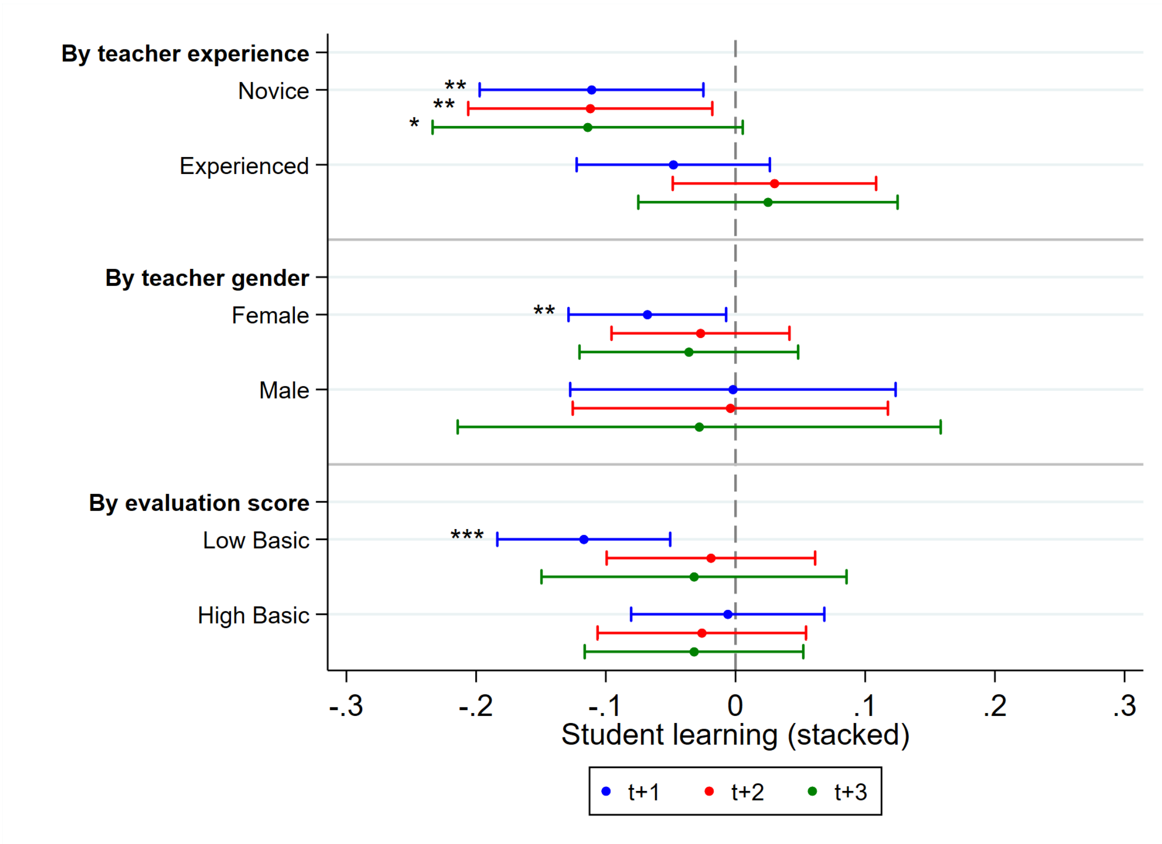

**(c)** Teacher re-evaluation scores and average student achievement




**(d)** Teacher re-evaluation score growth and average student achievement

*Note:* This figure examines the evaluation scores for re-evaluations among teachers who were assigned to re-evaluation and completed a re-evaluation. Panel (a) plots the distribution of re-evaluation scores. The rating categories resulting from this re-evaluation and the percentage of teachers in each category are printed within the score range for that category. Panel (b) examines the distribution of changes between initial evaluation scores and re-evaluation scores; higher values indicate improvements. These differences are standardized based on the standard deviation of evaluations from the initial evaluation year; the black vertical line plots the average change in scores in standard deviation units. Panels (c) and (d) are binned scatter plots of teacher re-evaluation scores and score growth against stacked student exam scores. Respectively, we either calculate ventiles of scores or ventiles of changes in scores and plot the average exam score within that bin. We also regress each exam score outcome on either teacher evaluation scores or score growth and plot the slope in red. In panels (a) and (c), the black dashed lines indicate the scores distinguishing the different rating categories; in panels (b) and (c), the black dashed line indicates no change in evaluation scores between the teacher's initial evaluation and the re-evaluation. In panel (b), the red dashed line indicates the expected regression to the mean for teachers below the re-evaluation threshold, calculated using the Mills ratio method with marginal reliability of 0.8 and accounting for conditional measurement error that increases in the tails—the observed change in scores significantly exceeds what can be explained by regression to the mean alone.

**Figure 2:** *Heterogeneity in effects on student learning by teacher characteristics*




*Note:* This figure plots coefficients and 95% confidence intervals from models examining heterogeneity in effects on stacked student SIMCE test scores in $t+1$, $t+2$, and $t+3$ based on different teacher characteristics. Models follow the same specification as Equation 1 but are subset by group (except for the bottom panel, where all observations are used, yet only low basic or high basic teachers are included as treated teachers, and the other basic teachers are dropped). Following previous literature on returns to teaching experience (Papay and Kraft, 2015; Bell et al., 2025) and previous research using the same 10-year threshold (Chetty et al., 2011), we construct the "novice" category to refer to teachers with less than or equal to 10 years of experience in the current year (whereas "experienced" refers to those with more than 10 years of experience). "Low basic" refers to teachers below the median basic score for evaluated teachers in the same year, and "high basic" to those above the median. *p<.1, **p<.05, ***p<.01

**Table 1:** *Compliance with re-evaluation law*

| | 2005-10 | | 2011-14 | | ITT |
|---|---|---|---|---|---|
| | Basic | Above | Basic | Above | effect |
| | (1) | (2) | (3) | (4) | (5) |
| **Panel A: Effect on teachers being re-evaluated** | | | | | |
| Re-evaluated in t+1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 (0.000) |
| *n* | 7,376 | 14,534 | 1,732 | 9,441 | 32,918 |
| Re-evaluated in t+2 | 0.000 | 0.000 | 0.588 | 0.000 | 0.589 (0.013)*** |
| *n* | 7,390 | 14,542 | 1,334 | 7,288 | 30,380 |
| Re-evaluated in t+3 | 0.000 | 0.000 | 0.002 | 0.000 | 0.002 (0.002) |
| *n* | 7,584 | 14,589 | 862 | 3,909 | 26,791 |

*Note:* This table reports on the law's effect on lower-performing ("basic") teachers being re-evaluated one, two, or three years after their initial evaluation (periods $t+1$, $t+2$, and $t+3$) based on whether their initial evaluation year was between 2005 and 2010 (pre-policy) or between 2011 to 2014 (when the policy was in effect). The law mandated that "basic" teachers be re-evaluated in period $t+2$. We calculate the mean re-evaluation rates for teachers rated basic or above in the pre- and post-policy periods in columns (1)-(4). In column (5) we report the coefficient on the interaction term between basic and the post-policy period (the difference-in-differences estimate) from a regression model predicting re-evaluation as the outcome that takes a similar form to Equation 1 except that it is run on the sample of teacher who have evaluations in the same $t+a$ period (for example, all previously evaluated teachers in their $t+2$ year). *p<.1, **p<.05, ***p<.01

**Table 2:** *Sample characteristics and validity checks*

| | 2005-10 | | 2011-14 | | ITT |
|---|---|---|---|---|---|
| | Basic | Above | Basic | Above | effect |
| | (1) | (2) | (3) | (4) | (5) |
| **Panel A: Lack of differential rating manipulation** | | | | | |
| Rating category different from score | 0.029 | 0.054 | 0.034 | 0.052 | 0.008 (0.006) |
| *n* | 7,376 | 14,534 | 1,732 | 9,441 | 32,918 |
| **Panel B: Absence of teacher sorting to/from standardized exams** | | | | | |
| Teaches SIMCE class in t+1 | 0.264 | 0.280 | 0.239 | 0.262 | -0.011 (0.013) |
| *n* | 7,376 | 14,534 | 1,732 | 9,441 | 32,918 |
| Teaches SIMCE class in t+2 | 0.254 | 0.272 | 0.221 | 0.252 | -0.013 (0.014) |
| *n* | 7,390 | 14,542 | 1,334 | 7,288 | 30,380 |
| Teaches SIMCE class in t+3 | 0.243 | 0.261 | 0.205 | 0.240 | -0.021 (0.017) |
| *n* | 7,584 | 14,589 | 862 | 3,909 | 26,791 |
| **Panel C: Balance of student characteristics** | | | | | |
| Female | 0.492 | 0.492 | 0.472 | 0.491 | -0.018 (0.014) |
| Attendance in t | 0.932 | 0.933 | 0.910 | 0.921 | -0.007 (0.002)*** |
| Grade point average (GPA) in t | 5.924 | 5.969 | 5.866 | 5.928 | -0.011 (0.017) |
| Repeated grade in t | 0.039 | 0.037 | 0.042 | 0.036 | 0.003 (0.004) |
| *n* | 62,997 | 159,244 | 8,882 | 69,854 | 734,549 |
| F statistic (p-value) | | | | | 0.397 |
| **Panel D: Balance of teacher characteristics** | | | | | |
| Female | 0.651 | 0.745 | 0.663 | 0.774 | -0.016 (0.016) |
| Contract hours | 39.017 | 38.561 | 39.262 | 38.930 | -0.470 (0.173)*** |
| Works in yet another school | 0.042 | 0.038 | 0.012 | 0.017 | -0.001 (0.005) |
| Years in service | 21.234 | 20.135 | 17.081 | 16.080 | -0.592 (0.397) |
| *n* | 2,295 | 5,044 | 410 | 2,505 | 54,271 |
| F statistic (p-value) | | | | | 0.660 |

*Note:* This table assesses threats to the validity of our identification strategy. It reports on whether the re-evaluation law affected score manipulation (Panel A), whether "basic" teachers taught classrooms whose students participated in the SIMCE exam (Panel B), and whether the law affected sorting based on student or teacher characteristics (Panels C and D). In columns (1)-(4), we calculate the mean of each covariate for teachers rated basic vs. above, based on whether their initial evaluation occurred between 2005-2010 (pre-policy) or 2011-2014 (post-policy). In column (5), we report the coefficient on the interaction term between basic and the post-policy period (the difference-in-differences estimate) from regression models that are variations of Equation 1. In Panel A, we calculate the proportion of teachers whose final evaluation rating does not match their continuous evaluation score (i.e., the rule-based evaluation rating was vetoed by the Municipal Evaluation Commission). In Panel B, we present the proportion of teachers teaching a fourth-grade classroom whose students participated in SIMCE. For panels A and B, we run versions of Equation 1 that subset to the sample of all primary school teachers who have evaluations in the same $t+a$ period (for example, all previously evaluated teachers in their $t+2$ year). Panels C and D examine the effects of the law on student and teacher characteristics in the $t+2$ re-evaluation year (results for $t+1$ and $t+3$ are in Appendix Table A1). Each of these characteristics is measured in baseline year $t$ prior to assignment to the current classroom (i.e., two years prior). For both panels, joint $F$-tests from regressions of the basic re-evaluation treatment on each of the covariates interacted with post-treatment indicators examine the null hypothesis that all of these coefficients are jointly equal to 0. The models for column (5) of Panel C take the same form as Equation 1, and for Panel D, they take the same form, but the models are run at the teacher level among the sample of teachers currently teaching a SIMCE class. *p<.1, **p<.05, ***p<.01

**Table 3:** *Intent-to-treat effects on student learning and teacher behaviors*

| | Year of assignment t+1 | Year of re-evaluation t+2 | Year after re-evaluation t+3 |
|---|---|---|---|
| | (1) | (2) | (3) |
| **Panel A: Student learning** | | | |
| Student learning (stacked) | -0.055** | -0.019 | -0.033 |
| | (0.028) | (0.031) | (0.039) |
| *n* | 798,609 | 734,549 | 548,005 |
| Language | -0.053* | -0.011 | 0.014 |
| | (0.030) | (0.033) | (0.039) |
| *n* | 401,201 | 370,187 | 277,564 |
| Mathematics | -0.062* | -0.025 | -0.081* |
| | (0.032) | (0.036) | (0.049) |
| *n* | 397,406 | 364,362 | 270,439 |
| **Panel B: Teachers** | | | |
| Caring | -0.055 | -0.000 | 0.037 |
| | (0.048) | (0.051) | (0.050) |
| *n* | 64,411 | 63,598 | 59,467 |
| Teaching practices | -0.031 | 0.048 | 0.021 |
| | (0.044) | (0.037) | (0.040) |
| *n* | 89,519 | 91,983 | 79,695 |

*Note:* This table examines intent-to-treat (ITT) impacts of being assigned to be re-evaluated on student learning and teacher behavior in the year of assignment to re-evaluation ($t+1$), the year of re-evaluation ($t+2$), and the year after re-evaluation ($t+3$). All models take the form of Equation 1. Panel A reports ITT effects on student learning. The study's main results rely on "stacking" test scores, allowing each student to enter the data twice, depending on whether they have a valid math and reading score, and because they may have different math and language teachers on different evaluation timelines. We also report on language and math scores separately based on the evaluation scores of the language or math teacher, respectively (for 80% of students, this is the same teacher). Panel B reports ITT effects on parent- and student-reported teacher behaviors. Since these surveys ask about the head teacher or alternate between asking about math and language teachers, we restrict these analyses to only those students who have the same math and language teacher. For caring, parents were asked how caring the head teacher was on a 7-point scale; we standardized ratings within each year. This measure was available from 2011 to 2014. For teaching practices, students were asked six questions about what their teachers did in the classroom, which we combined into an index using Item Response Theory (IRT), standardizing scores within each year. This measure was available from 2011 to 2015. For both measures, higher values indicate more favorable ratings of teacher behaviors. *p<.1, **p<.05, ***p<.01

## References

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

**Aucejo, Esteban, Teresa Romano, and Eric S. Taylor.** 2022. "Does Evaluation Change Teacher Effort and Performance? Quasi-experimental Evidence from a Policy of Retesting Students." *The Review of Economics and Statistics*, 104(3): 417–430.

**Baker, George P.** 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy*, 100(3): 598–614.

**Barlevy, Gadi, and Derek Neal.** 2012. "Pay for Percentile." *The American Economic Review*, 102(5): 1805–1831.

**Barrios-Fernández, Andrés, and Marc Riudavets-Barcons.** 2024. "Teacher Value-added and Gender Gaps in Educational Outcomes." *Economics of Education Review*, 100.

**Bell, Courtney, Jessalynn James, Eric S. Taylor, and James Wyckoff.** 2025. "Measuring Returns to Experience Using Supervisor Ratings of Observed Performance: The Case of Classroom Teachers." *Journal of Policy Analysis and Management*, 44(1): 12–44.

**Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer.** 2025. "Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment." *Journal of Political Economy Microeconomics*, 3(3): 568–610.

**Briole, Simon, and Eric Maurin.** 2024. "There's Always Room for Improvement: The Persistent Benefits of a Large-Scale Teacher Evaluation System." *Journal of Human Resources*, 59(4): 1150–1179.

**Burgess, Simon, Shenila Rawal, and Eric S. Taylor.** 2021. "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools." *Journal of Labor Economics*, 39(4): 1155–1186.

**Buurman, Margaretha, Josse Delfgaauw, Robert Dur, and Robin Zoutenbier.** 2020. "When Do Teachers Respond to Student Feedback? Evidence From a Field Experiment." *Labour Economics*, 65: 101858.

**Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan.** 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics*, 126(4): 1593–1660.

**Dee, Thomas S., and James Wyckoff.** 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT: Incentives, Selection, and Teacher Performance." *Journal of Policy Analysis and Management*, 34(2): 267–297.

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

**Gardner, John.** 2022. "Two-Stage Differences in Differences." arXiv:2207.05943 [econ].

**Goldhaber, Dan, and Emily Anthony.** 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *The Review of Economics and Statistics*, 89(1): 134–150.

**Holmström, Bengt, and Paul Milgrom.** 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *The Journal of Law, Economics, and Organization*, 7(Special Issue): 24–52.

**Jackson, C. Kirabo.** 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on NonTest Score Outcomes." *Journal of Political Economy*, 126(5): 2072–2107.

**Jacob, Brian.** 2011. "Do Principals Fire the Worst Teachers?" *Educational Evaluation and Policy Analysis*, 33(4).

**Jacob, Brian, and Jesse Rothstein.** 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives*, 30(3): 85–108.

**Koedel, Cory, Jiaxi Li, Matthew G. Springer, and Li Tan.** 2017. "The Impact of Performance Ratings on Job Satisfaction for Public School Teachers." *American Educational Research Journal*, 54(2): 241–278.

**Koedel, Cory, Jiaxi Li, Matthew G. Springer, and Li Tan.** 2019. "Teacher Performance Ratings and Professional Improvement." *Journal of Research on Educational Effectiveness*, 12(1): 90–115.

**Kraft, Matthew A.** 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher*, 49(4): 241–253.

**Loeb, Susanna, Luke Miller, and James Wyckoff.** 2015. "Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City." *Educational Researcher*, 44(4).

**Papay, John P., and Matthew A. Kraft.** 2015. "Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement." *Journal of Public Economics*, 130: 105–119.

**Sartain, Lauren, and Matthew P. Steinberg.** 2016. "Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools." *Journal of Human Resources*, 51(3): 615–655.

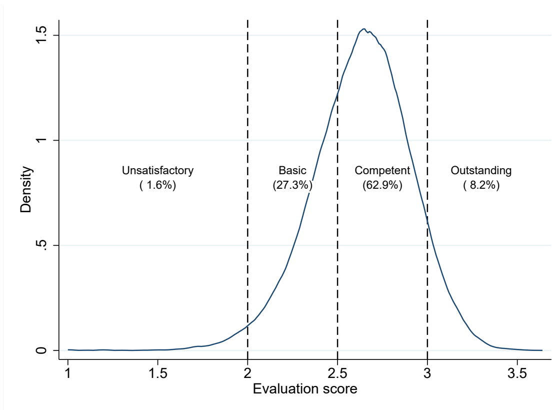
**Taylor, Eric S., and John H. Tyler.** 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review*, 102(7): 3628–3651.

**The World Bank.** 2025. "Education Statistics (EdStats)."

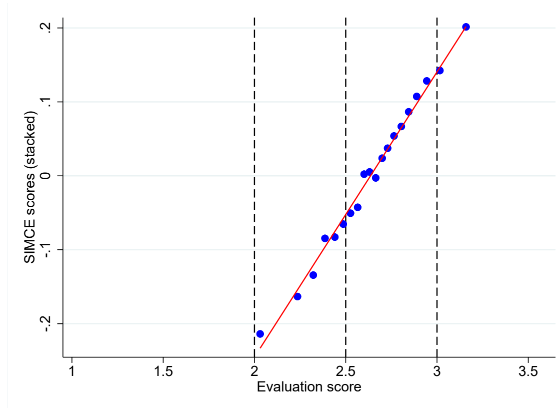
# Appendix

## A Additional figures and tables

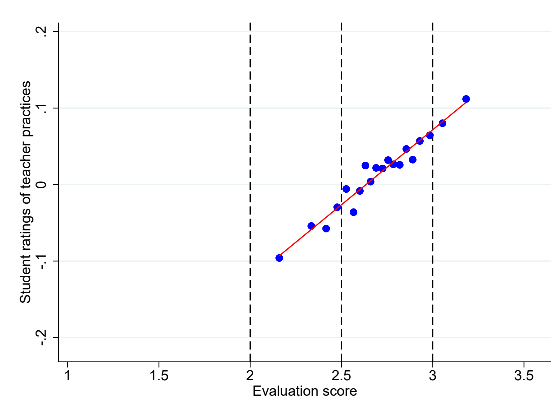
**Figure A1:** *Distribution of teacher evaluation scores and their associations with outcomes*



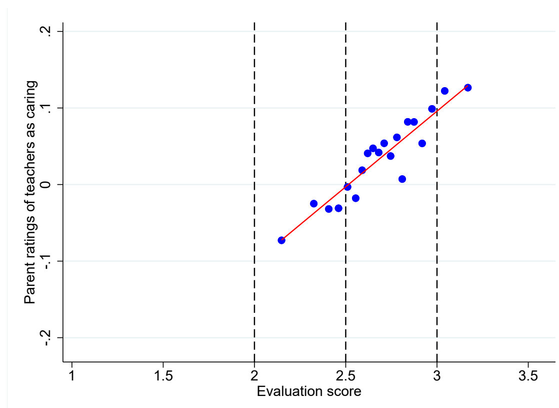

**(a)** Distribution of teacher evaluation scores




**(b)** Teacher evaluation scores and average student achievement



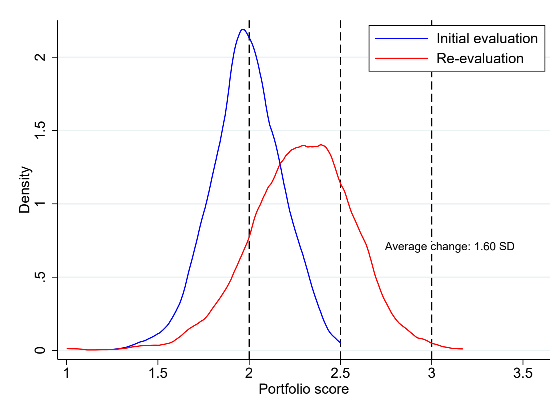

**(c)** Teacher evaluation scores and student ratings of teaching practices



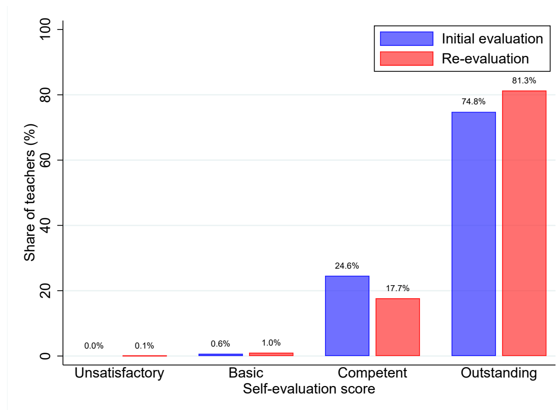

**(d)** Teacher evaluation scores and caregiver ratings of teacher caring

*Note:* This figure plots the distribution in teacher evaluation scores and their relationship to student achievement and teacher behavior. Throughout, teacher evaluation scores are on the x-axis with dashed vertical black lines indicating the scores that distinguish the overall rating levels. Specifically, panel (a) plots the distribution of scores on their evaluations (but not any re-evaluations) for all teachers in the sample. The rating categories and the percentage of teachers in that category are printed within the score range for that category. Panels (b) - (d) are binned scatter plots of teacher evaluation scores against outcomes. For each panel, we calculate ventiles of teacher evaluation scores for the relevant years and calculate the average outcome within that bin. We also regress each outcome on teacher evaluation scores and plot the slope in red. Panel (b) plots average standardized achievement pooled across language and math for the students taught by that teacher (years 2005-2015). Panel (c) plots average standardized scores from an index of six questions asking students about the teaching practices their teacher engages in that we combine using Item Response Theory (years 2011-2015). Panel (d) plots the average standardized value that teachers receive from ratings by parents of the students they teach of their level of caring (years 2011-2014).

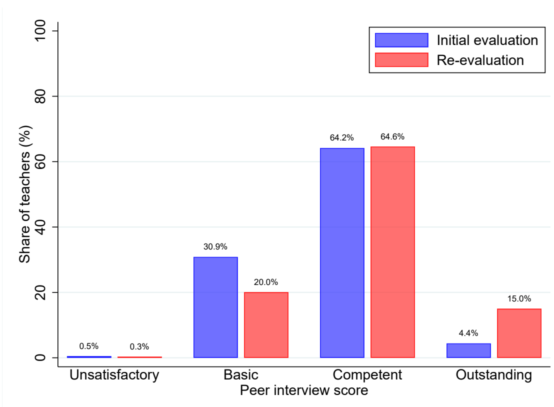
**Figure A2:** *Changes in score by domain from initial evaluation to re-evaluation*



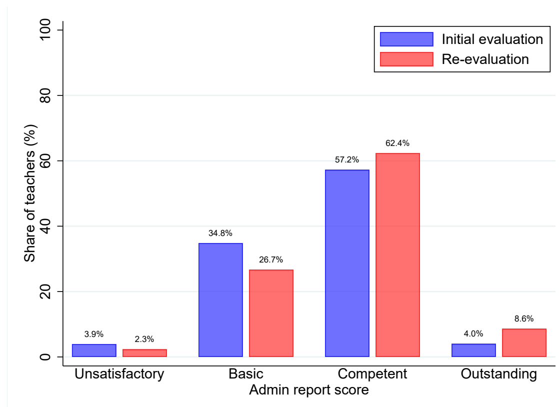

**(a)** Portfolio scores (60%)




**(b)** Self-evaluation scores (10%)



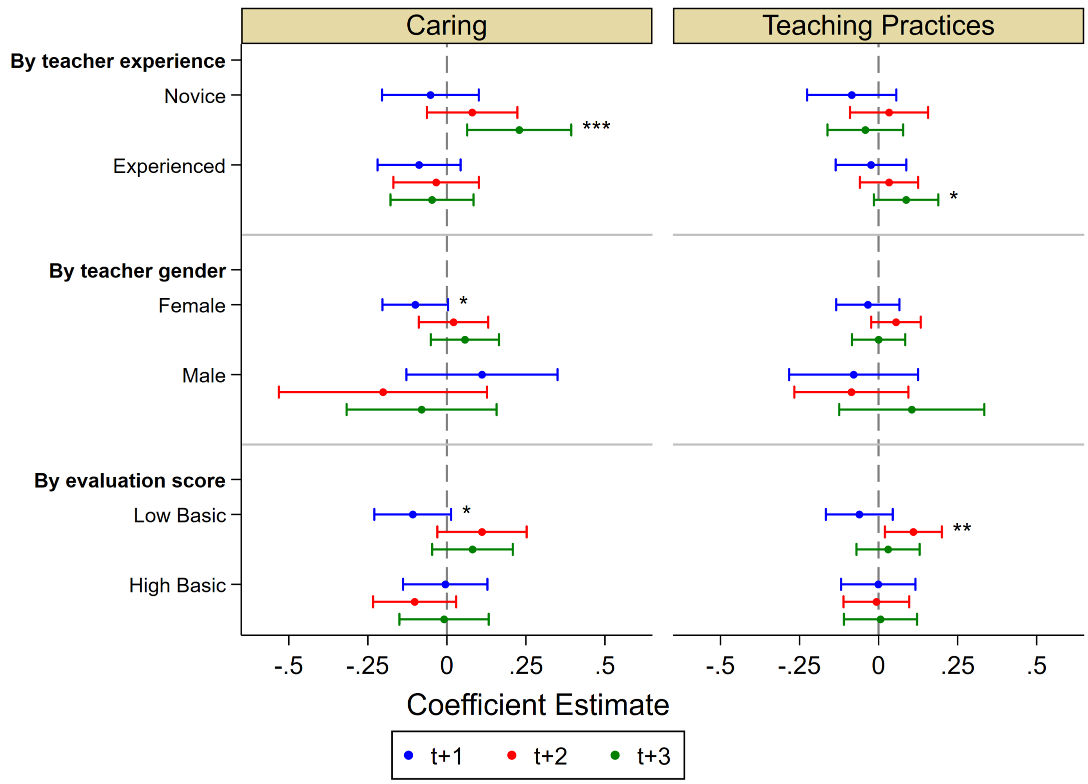

**(c)** Peer interview scores (20%)




**(d)** Administrator report scores (10%)

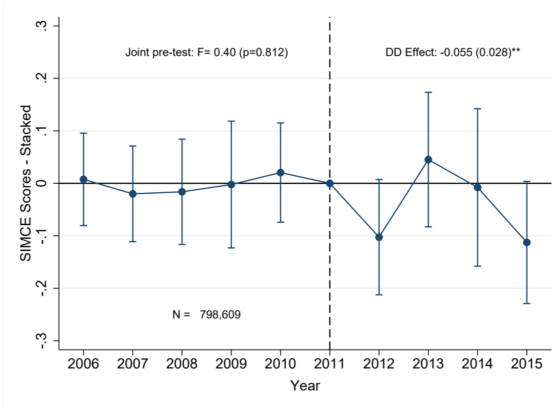
*Note:* This figure examines the evaluation scores by domain for the initial evaluation and re-evaluation among the teachers that were re-evaluated. Next to each domain is the percent that the domain is weighted for the total final score. Panel (a) is the continuous portfolio score. The difference between the initial and re-evaluation score is standardized based on the standard deviation of evaluations from the initial evaluation year; the standard deviation change is printed on the figure. The black dashed lines represent the cutoffs for each category. Panels (b)-(d) plot the share of re-evaluated teachers in each of the categories for the initial evaluation and re-evaluation.

**Figure A3:** *Heterogeneity in effects on teacher behavior by teacher characteristics*

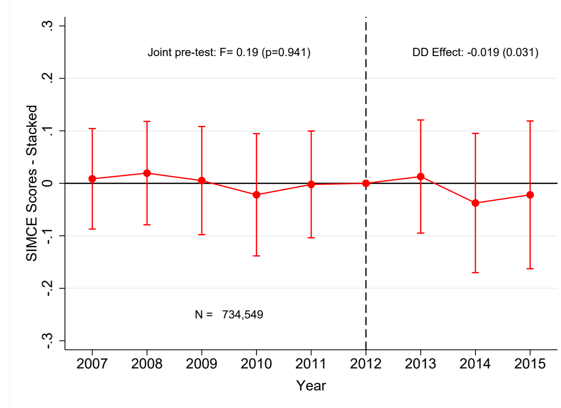



*Note:* This figure plots coefficients and 95% confidence intervals from models examining heterogeneity in effects on teacher behaviors in $t+1$, $t+2$, and $t+3$ based on different teacher characteristics. Models follow the same specification as Equation 1 but are subset by group (except for the bottom panel, where all observations are used, yet only low basic or high basic teachers are included as treated teachers, and the other basic teachers are dropped). Following previous literature on returns to teaching experience (Papay and Kraft, 2015; Bell et al., 2025) and previous research using the same 10-year threshold (Chetty et al., 2011), we construct the "novice" category to refer to teachers with less than or equal to 10 years of experience in the current year (whereas "experienced" refers to those with more than 10 years of experience). "Low basic" refers to teachers below the median basic score for evaluated teachers in the same year, and "high basic" to those above the median. *p<.1, **p<.05, ***p<.01

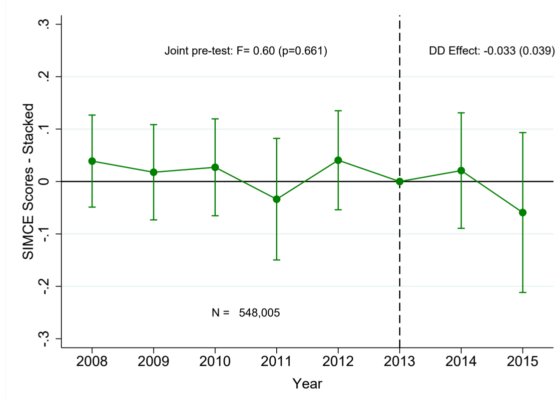
**Figure A4:** *Event studies for stacked SIMCE test scores.*




**(a)** *t+1*



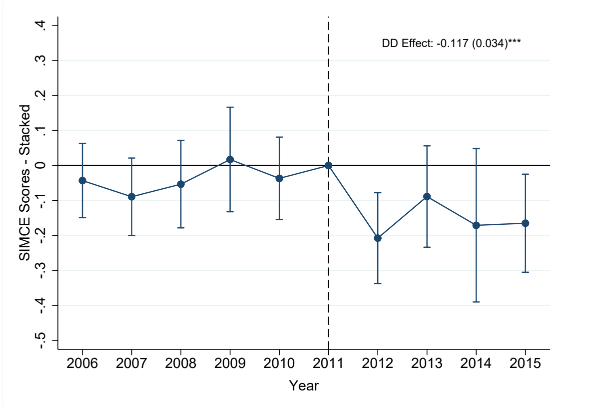

**(b)** *t+2*



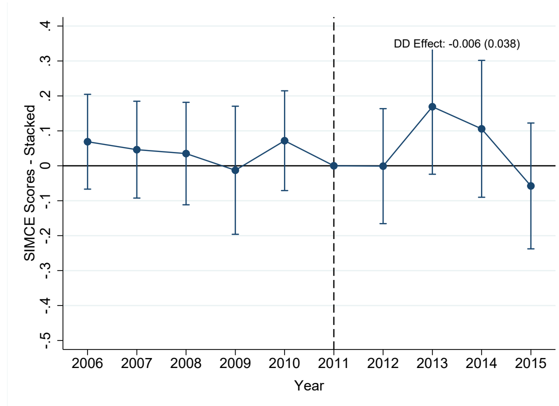

**(c)** *t+3*

*Note:* These figures are event studies for the ITT impacts of a basic rating on stacked student learning in the $t+1$, $t+2$, and $t+3$ periods relative to the omitted year. For each panel, models are estimated using Equation 1, except that the post-treatment indicator, interacted with the basic rating, is replaced with year indicators to examine trends in the outcome before and after policy implementation. For each model, we omit the last year indicator prior to when the re-evaluation policy would start impacting evaluated teachers (for example, in panel (a), we omit 2011 because teachers one year after their initial evaluation in 2011 would have initially been evaluated in 2010 and not subject to the policy). Black dashed lines indicate when teachers begin being impacted by the policy. In each panel, we test whether the pre-treatment coefficients are jointly different from 0 and report the $F$-statistic and $p$-value. We also report the static difference-in-differences estimates as presented in Table 3.

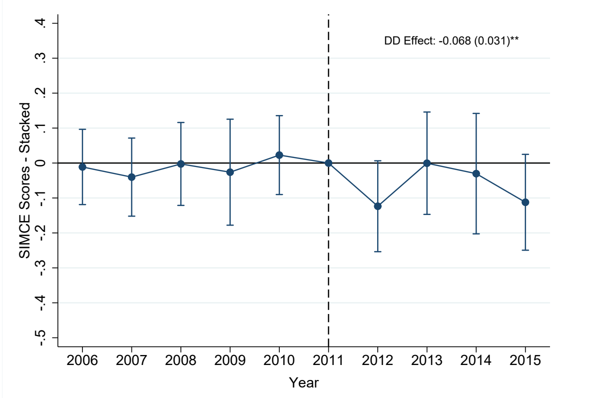
**Figure A5:** *Event studies for heterogeneous effects - Evaluation score and gender*



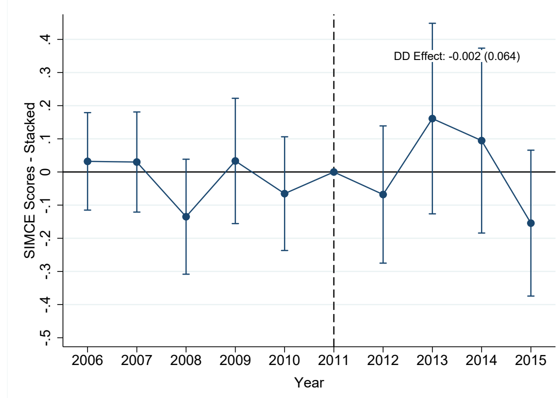

**(a)** *t* + *1* - Low Basic




**(b)** *t* + *1* - High Basic



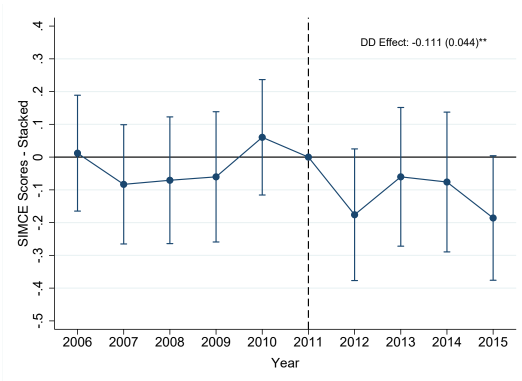

**(c)** *t* + *1* - Female



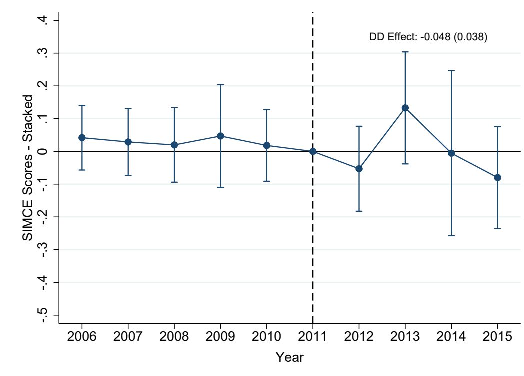

**(d)** *t* + *1* - Male

*Note:* These figures are event studies for the heterogeneous ITT impacts of a basic rating on stacked student learning relative to the omitted year. We produce event studies only for the sub-groups and years where we detect significant effects for one of the groups (see Figure 2). Panels (a) and (b) present the $t+1$ results for low and high basic teachers, whereas panels (c) and (d) do so for female and male teachers. For each panel, models are estimated following Equation 1 and as described in Figure 2, except that the post-treatment indicator interacted with the basic rating is replaced with year indicators. For each model, we omit the last year indicator prior to when the re-evaluation policy would start impacting evaluated teachers (for example, in panel (a), we omit 2011 because teachers one year after their initial evaluation in 2011 would have initially been evaluated in 2010 and not subject to the policy). Black dashed lines indicate when teachers begin being impacted by the policy. We also report the static difference-in-differences estimates as presented in Figure 2.

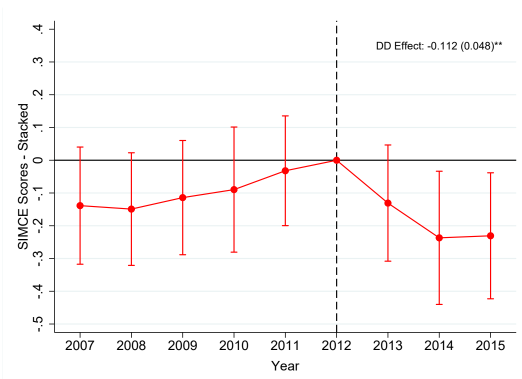
**Figure A6:** *Event studies for heterogeneous effects - Teacher experience*



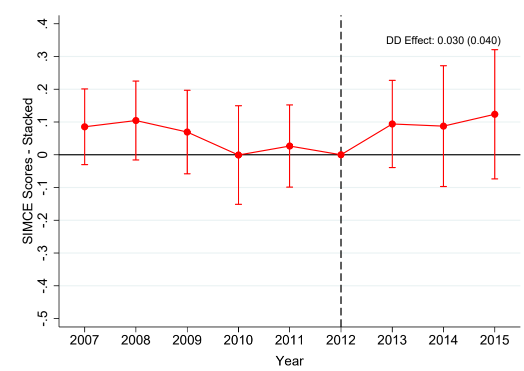

**(a)** $t + 1$ - Novice




**(b)** $t + 1$ - Experienced

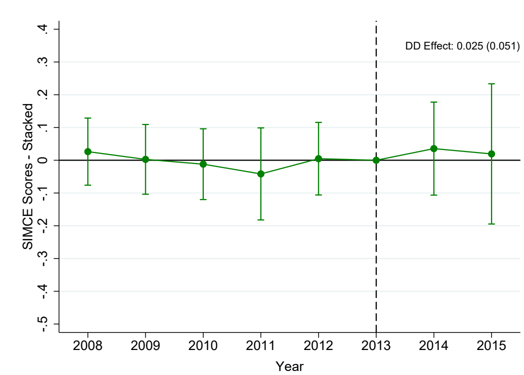



**(c)** $t + 2$ - Novice




**(d)** $t + 2$ - Experienced


**(e)** $t + 3$ - Novice




**(f)** $t + 3$ - Experienced

*Note:* These figures are event studies for the heterogeneous ITT impacts of a basic rating on stacked student learning relative to the omitted year. We produce event studies only for the sub-groups and years where we detect significant effects for one of the groups (see Figure 2). Here we present results for novice ($<=$ 10 years experience) and experienced ($>$ 10 years) for $t+1$, $t+2$, and $t+3$. For each panel, models are estimated following Equation 1 and as described in Figure 2 except that the post-treatment indicator interacted with the basic rating is swapped out for year indicators. For each model, we omit the last year indicator prior to when the re-evaluation policy would start impacting evaluated teachers. Black dashed lines indicate when teachers begin being impacted by the policy. We also report the static difference-in-differences estimates as presented in Figure 2.

**Table A1:** *Sample characteristics and validity checks (t+1 and t+3)*

| | 2005-10 | | 2011-14 | | ITT |
|---|---|---|---|---|---|
| | Basic | Above | Basic | Above | effect |
| | (1) | (2) | (3) | (4) | (5) |
| **Panel A: Balance of teacher characteristics (t+1)** | | | | | |
| Gender: Female | 0.640 | 0.735 | 0.676 | 0.762 | -0.018 (0.014) |
| Contract hours | 39.018 | 38.523 | 38.826 | 38.823 | -0.410 (0.159)** |
| Works in yet another school | 0.049 | 0.038 | 0.018 | 0.021 | 0.001 (0.005) |
| Years in service | 20.508 | 19.570 | 15.122 | 14.863 | -0.292 (0.334) |
| *n* | 2,314 | 5,046 | 558 | 3,345 | 55,399 |
| F statistic (p-value) | | | | | 0.126 |
| **Panel B: Balance of student characteristics (t+1)** | | | | | |
| Gender: Female | 0.481 | 0.488 | 0.480 | 0.491 | -0.003 (0.011) |
| Attendance in t | 0.933 | 0.934 | 0.913 | 0.922 | -0.005 (0.002)*** |
| GPA in t | 5.798 | 5.850 | 5.778 | 5.827 | 0.002 (0.017) |
| Repeated grade in t | 0.033 | 0.027 | 0.027 | 0.022 | 0.001 (0.003) |
| *n* | 68,541 | 164,013 | 12,113 | 91,247 | 798,609 |
| F statistic (p-value) | | | | | 0.030 |
| **Panel C: Balance of teacher characteristics (t+3)** | | | | | |
| Gender: Female | 0.659 | 0.756 | 0.717 | 0.777 | -0.002 (0.019) |
| Contract hours | 39.174 | 38.858 | 39.206 | 39.276 | -0.508 (0.200)** |
| Works in yet another school | 0.036 | 0.036 | 0.013 | 0.015 | -0.003 (0.006) |
| Years in service | 22.638 | 21.218 | 16.756 | 16.953 | -0.763 (0.485) |
| *n* | 2,310 | 4,964 | 240 | 1,354 | 47,807 |
| F statistic (p-value) | | | | | 0.225 |
| **Panel D: Balance of student characteristics (t+3)** | | | | | |
| Gender: Female | 0.483 | 0.488 | 0.462 | 0.494 | -0.026 (0.014)* |
| Attendance in t | 0.928 | 0.929 | 0.911 | 0.919 | -0.002 (0.003) |
| GPA in t | 6.042 | 6.076 | 6.028 | 6.045 | 0.001 (0.023) |
| Repeated grade in t | 0.044 | 0.042 | 0.045 | 0.044 | -0.002 (0.005) |
| *n* | 64,663 | 151,811 | 5,402 | 37,225 | 548,005 |
| F statistic (p-value) | | | | | 0.063 |

*Note:* This table assesses threats to the validity of the design by reporting on whether the re-evaluation law triggered endogenous sorting based on student or teacher characteristics. We calculate the mean rates of each covariate for teachers rated basic or above based on whether their initial evaluation occurred between 2005-2010 (pre-policy) or between 2011-2014 (post-policy) in columns (1)-(4). In column (5), we report the coefficient on the interaction term between basic and the post-policy period (the DiD estimate) from regression models that are variations of equation 1. Panels A and B examine covariates for teacher and student characteristics in $t+1$ and panels C and D do so for $t+3$ (see Table 2 for $t+2$ results). Each of these characteristics is measured in baseline year $t$ prior to assignment to the current classroom (i.e., one or three years prior). For all panels, joint $F$-tests from regressions of the basic re-evaluation treatment on each of the covariates interacted with post-treatment indicators examine the null hypothesis that all of these coefficients are jointly equal to 0. The models for column (5) of panels B and D follow the same form as Equation 1. For panels A and C, the models take the same form, but they are run at the teacher level among the sample of teachers currently teaching a SIMCE class. *p<.1, **p<.05, ***p<.01

**Table A2:** *Robustness checks*

| | Falsification | With student baseline controls | | | Two-Stage DiD (Gardner, 2022) | | |
|---|---|---|---|---|---|---|---|
| | Year before evaluation t-1 | Year of assignment t+1 | Year of re-evaluation t+2 | Year after re-evaluation t+3 | Year of assignment t+1 | Year of re-evaluation t+2 | Year after re-evaluation t+3 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| **Panel A: Student learning** | | | | | | | |
| Student learning (stacked) | -0.035 | -0.059** | -0.015 | -0.034 | -0.054* | -0.020 | -0.031 |
| | (0.040) | (0.028) | (0.030) | (0.036) | (0.028) | (0.031) | (0.040) |
| *n* | 158,138 | 790,679 | 720,983 | 534,723 | 798,609 | 734,549 | 548,005 |
| Language | -0.044 | -0.064** | -0.013 | 0.013 | -0.052* | -0.011 | 0.015 |
| | (0.044) | (0.028) | (0.033) | (0.035) | (0.030) | (0.034) | (0.039) |
| *n* | 78,915 | 397,226 | 363,300 | 270,877 | 401,202 | 370,187 | 277,565 |
| Mathematics | -0.024 | -0.062* | -0.014 | -0.080* | -0.061* | -0.028 | -0.076 |
| | (0.045) | (0.034) | (0.036) | (0.045) | (0.033) | (0.037) | (0.051) |
| *n* | 79,223 | 393,451 | 357,683 | 263,846 | 397,407 | 364,362 | 270,440 |
| **Panel B: Teacher behavior** | | | | | | | |
| Caring | | -0.060 | -0.004 | 0.041 | -0.054 | 0.003 | 0.034 |
| | | (0.048) | (0.052) | (0.050) | (0.049) | (0.053) | (0.053) |
| *n* | | 63,770 | 62,696 | 58,244 | 64,414 | 63,600 | 59,468 |
| Teaching practices | | -0.036 | 0.048 | 0.025 | -0.033 | 0.049 | 0.024 |
| | | (0.043) | (0.037) | (0.040) | (0.043) | (0.038) | (0.044) |
| *n* | | 88,415 | 90,373 | 77,752 | 89,520 | 91,983 | 79,696 |

*Note:* This table examines the robustness of the intent-to-treat (ITT) impacts of those assigned an initial basic rating on student learning and teacher behavior in the year of assignment to re-evaluation ($t+1$), the year of re-evaluation ($t+2$), and the year after re-evaluation ($t+3$). In column (1), we perform a falsification test where we examine the impact of being rated basic in the *next* school year on the test scores of students taught by the eventually treated teacher (we can only do this for test scores due to the years of data available for the teacher behavior outcomes). In columns (2)-(4), we estimate the same models as in Table 3 but add in controls for student gender as well as GPA, attendance, and grade repetition in baseline year $t$. In columns (5)-(7), we present models that take the same general form as in Table 3 but are estimated using the Gardner (2022) two-stage DiD approach that is robust to well-known negative weighting problems that can occur in contexts with staggered treatment timing and potentially heterogeneous treatment effects. *p<.1, **p<.05, ***p<.01

**Table A3:** *Heterogeneity in effects by teacher characteristics*

| | Student Learning (Stacked) | | | Caring | | | Teaching practices | | |
|---|---|---|---|---|---|---|---|---|---|
| | Post-assignment (t+1) | Re-evaluation (t+2) | Post re-evaluation (t+3) | Post-assignment (t+1) | Re-evaluation (t+2) | Post re-evaluation (t+3) | Post-assignment (t+1) | Re-evaluation (t+2) | Post re-evaluation (t+3) |
| **Panel A: By teacher experience** | | | | | | | | | |
| Novice (<=10 years) | -0.111** | -0.112** | -0.114* | -0.052 | 0.080 | 0.229*** | -0.085 | 0.033 | -0.042 |
| | (0.044) | (0.048) | (0.061) | (0.078) | (0.073) | (0.084) | (0.072) | (0.063) | (0.061) |
| Experienced (>10 years) | -0.048 | 0.030 | 0.025 | -0.088 | -0.034 | -0.047 | -0.024 | 0.033 | 0.087* |
| | (0.038) | (0.040) | (0.051) | (0.067) | (0.069) | (0.067) | (0.057) | (0.047) | (0.052) |
| Difference | -0.034 | -0.097 | -0.135* | 0.036 | 0.122 | 0.182* | -0.055 | 0.009 | -0.121 |
| | (0.055) | (0.059) | (0.075) | (0.094) | (0.100) | (0.094) | (0.086) | (0.074) | (0.077) |
| *n* | 798,609 | 734,549 | 548,005 | 64,411 | 63,598 | 59,467 | 89,519 | 91,983 | 79,695 |
| **Panel B: By teacher evaluation score** | | | | | | | | | |
| Basic-low | -0.117*** | -0.019 | -0.032 | -0.108* | 0.111 | 0.081 | -0.061 | 0.110** | 0.030 |
| | (0.034) | (0.041) | (0.060) | (0.062) | (0.072) | (0.065) | (0.054) | (0.046) | (0.051) |
| Basic-high | -0.006 | -0.026 | -0.032 | -0.005 | -0.102 | -0.009 | -0.001 | -0.007 | 0.006 |
| | (0.038) | (0.041) | (0.043) | (0.068) | (0.067) | (0.072) | (0.060) | (0.053) | (0.059) |
| Difference | 0.105 | -0.032 | -0.031 | -0.008 | -0.104* | -0.021 | -0.039 | 0.038 | 0.010 |
| | (0.081) | (0.088) | (0.100) | (0.064) | (0.060) | (0.066) | (0.057) | (0.047) | (0.052) |
| *n* | 798,609 | 734,549 | 548,005 | 64,411 | 63,598 | 59,467 | 89,519 | 91,983 | 79,695 |
| **Panel C: By teacher gender** | | | | | | | | | |
| Female | -0.068** | -0.027 | -0.036 | -0.100* | 0.021 | 0.057 | -0.034 | 0.055 | -0.000 |
| | (0.031) | (0.035) | (0.043) | (0.053) | (0.056) | (0.055) | (0.051) | (0.040) | (0.043) |
| Male | -0.002 | -0.004 | -0.028 | 0.111 | -0.202 | -0.080 | -0.079 | -0.086 | 0.105 |
| | (0.064) | (0.062) | (0.095) | (0.122) | (0.168) | (0.121) | (0.104) | (0.092) | (0.117) |
| Difference | -0.028 | -0.028 | 0.053 | -0.131 | 0.085 | 0.089 | -0.021 | 0.040 | -0.115 |
| | (0.068) | (0.062) | (0.093) | (0.120) | (0.130) | (0.136) | (0.091) | (0.095) | (0.112) |
| *n* | 798,609 | 734,549 | 548,005 | 64,411 | 63,598 | 59,467 | 89,519 | 91,983 | 79,695 |

*Note:* The table reports coefficients and 95% confidence intervals from models examining heterogeneity in effects on student learning and teacher behavior in $t+1$, $t+2$ and $t+3$ based on different teacher characteristics. These coefficients match what is reported in Figure 2. Models follow the same specification as Equation 1 but are subset by group (except for the bottom panel, where all observations are used, except that only low basic or high basic teachers are included as treated teachers and the other basic teachers are dropped). We formally test the differences between the sub-group coefficients through three-way interactions with basic, a post-treatment indicator, and a dummy for one of the sub-groups. "Low basic" refers to teachers below the median basic score for evaluated teachers in the same year, and "high basic" to those above the median. *p<.1, **p<.05, ***p<.01