Evaluating Teacher Evaluation:
Evidence from Chile

Andreas de Barros[☆]

Abstract

This study investigates the causal effects of repeat, formative performance evaluations, under Chile's national teacher evaluation system. The study's main results suggest that student learning, teacher beliefs and teaching behaviors are not positively affected when evaluations are mandated, both in the year of the evaluation and in the year thereafter. These findings rest on data-sources with unusually comprehensive coverage of a national education system—positive effects on student performance are thus ruled out precisely. The results are not driven by a teacher's level of work experience, by student sorting, by systematic attrition, or by the article's model specification.

Highlights

- Investigates the causal effects of repeat, formative teacher evaluations.
- First paper to study effects on student learning under a national system.
- Rules out positive effects on learning, especially for novice teachers.
- Finds both null and negative effects on teachers' behavior and beliefs.

JEL codes: I20; I21; I28; J24.

Keywords: Education; employee productivity; formative evaluations; human capital; performance evaluations; teacher evaluations.

[☆]*Andreas de Barros is a PhD Candidate at Harvard University, 13 Appian Way, Cambridge, MA 02138.* *For support and comments, the author would like to thank 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, Abhijeet Singh, Ugo Troiano, and Martin West. The author thanks the Chilean Agencia de Calidad de la Educación and the Ministry of Education for making data available. The usual disclaimer applies. This paper uses confidential data maintained by the Chilean Ministry of Education. The data can be obtained by filing a request directly with the Ministry (see http://formulario.agenciaeducacion.cl/solicitud_cargar). The author is willing to assist (adebarros@g.harvard.edu). All programs and code will be made available as an Online Appendix. The author has nothing to disclose.*

1. Introduction

Performance evaluations are among the most controversial attempts to improve teacher effectiveness, especially if they are based on student test score gains and if they are used to inform teacher compensation and dismissal.[1] In contrast, there are frequent demands for more comprehensive, "formative" evaluation systems (see Grissom and Youngs, 2016). In its call for such a formative approach, for example, the United States' largest teacher's labor union defines the core purpose of teacher assessment and evaluation as to "strengthen the knowledge, skills, dispositions, and classroom practices of professional educators" (as opposed to a "rewards-and-punishment framework") (National Education Association, 2019, 1).

Proponents of such formative evaluations often refer to Chile's national evaluation system as a best-practice example. Chile's system embraces core principles of a formative teacher evaluation approach, such as the promotion of professional development, open collaboration and transparency, the use of multiple measures, validation, links to clear teaching standards (based on a "Framework of Good Teaching"), and the system's co-creation with teachers. A recent World Bank report therefore concludes that, while "[p]utting in place a sound system of teacher evaluation is expensive and institutionally challenging", "Chile's comprehensive teacher evaluation system, Docentemas [sic], has shown that it can be done" (Bruns and Luque, 2014, 215 et sq.).[2]

This study evaluates the causal effects of repeat, formative performance evaluations—under Chile's national teacher evaluation system "Docentemás"[3]. The article answers

[1] For recent reviews of such evaluation policies, see Jackson and Cowan (2018) and Lovison and Taylor (2018).

[2] The same report concludes that Chile's teacher evaluation system "remains the [Latin American] region's best practice example to date" (*ibid.*, 35).

[3] Formally, Docentemás is called the "Sistema de Evaluación del Desempeño Profesional Docente". Commonly, it is also referred to as "Evaluación Docente".

three main questions. First and foremost, do these formative evaluations lead to increased teacher effectiveness, as measured by student learning? Second, how are potential mechanisms affected that are expected to enhance student learning? Formative evaluations seek to improve instructional practices and to alter commonly held beliefs among teachers—I therefore investigate impacts on these intermediary factors. Third, do evaluations affect less-experienced teachers more strongly? Previous research on the returns to teacher experience and on teachers' dynamic skill development suggests that productivity improvements predominantly occur during the first five to ten years on the job.[4] Hence, I assess heterogeneous effects of evaluations by teachers' level of work experience.

A key challenge for the estimation of causal effects of teacher evaluations is the endogeneity of a teacher's assignment to evaluations.[5] To overcome this challenge, this study's analytical strategy relies on a difference-in-difference estimation strategy. More specifically, I exploit a policy change in the assignment mechanism. In 2011, Chile passed a new law, requiring teachers ranked in the "basic" (the second lowest) performance category to be re-evaluated after two years (instead of four). My identification strategy leverages this variation across time. I show that, although the law is imperfectly observed, it sharply increased a "basic" teacher's likelihood of being newly evaluated after two years. In additional analyses, the article moreover confirms that common assumptions of a difference-in-difference estimator are met, and that its analyses are not compromised by systematic student sorting or by differential attrition.

[4]For a recent overview, see Kraft et al. (2020).

[5]The direction of this bias is unclear. Formative evaluations may be assigned to weaker teacher of greatest need of personal development. Yet, more motivated, stronger teachers may also self-select into formative evaluations as they seek out personal development opportunities.

These analyses rest on data sources with unusually comprehensive coverage of a national education system. For the years 2005 to 2015, I use teacher-classroom links to match data on the universe of elementary teachers in Chile's public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students (covering more than 30 million student-by-year observations), and results on standardized test scores for all Chilean fourth graders in mathematics and language. I further complement these data with information on teaching and teacher behaviors from teacher surveys, student surveys, and parent surveys.

The study's main results suggest that student learning remains unaffected by a teacher's requirement to undergo a formative evaluation, both in the year of the evaluation and in the year thereafter. Intent-to-treat effects are precisely estimated, ruling out positive impacts of 0.03 and 0.08 standard deviations, respectively. In analyses of potential mechanisms, the study documents how teacher beliefs and teaching behaviors also remain unimproved. If anything, the findings point to detrimental effects of formative teacher evaluations. These results do not differ for teachers with fewer years of work experience. In robustness checks, I moreover show how they do not depend on model specifications and that their qualitative conclusions remain unaltered if a slightly modified assignment definition is used.

These analyses and their findings are novel in three distinct ways. They represent only the second causal investigation on the impact of formative teacher evaluations on student learning (and the first to include additional analyses of potential mediators). They moreover offer first evidence on the impact of repeat evaluations (i.e., the effect of *regularly* subjecting teachers to evaluations). To my best knowledge, this study furthermore provides the first quasi-experimental assessment of a national workforce evaluation system's effects on worker productivity—hence, it may also offer interesting insights for

public human resource management beyond the education sector.

The study thus contributes to several strands of a nascent literature within the economics of education and public economics, on the effectiveness of formative performance evaluations. One related body of literature has studied the effects of teacher evaluations on other teacher-level outcomes, such as professional improvement activities (Koedel et al., 2019), effort (Aucejo et al., 2019), job satisfaction (Koedel et al., 2017), and labor market responses (Sartain and Steinberg, 2016). Another collection of studies has focused on the effect of sub-components that may be part of formative evaluation systems, including in-person classroom observations (Burgess et al., 2019; Kane et al., 2019), peer collaboration (such as lesson study and instructional rounds) (Gersten et al., 2010; Louis and Marks, 1998), tutoring, mentoring, and coaching (Allen et al., 2011; Papay et al., 2019; Kraft et al., 2018; Kraft and Hill, 2019), the provision of formative feedback (Garet et al., 2017), and the release of teacher performance scores (Bergman and Hill, 2018; Pope, 2019).

Strikingly, however, there is so far only one other study on the causal effects of formative teacher evaluations on student performance.[6] For a sample of 105 mid-career teachers in Cincinnati Public Schools, Taylor and Tyler (2012) apply a teacher fixed-effects strategy. They find that math test scores of students whose teacher was evaluated in the previous year increased by about ten percent of a standard deviation. Taylor and Tyler (2012) also conclude that this effect is greater for teachers whose previous performance was lower. They cannot reject the absence of effects on reading scores.

[6] Steinberg and Sartain (2015) study the effects of a formative teacher evaluation pilot program on *school* performance. In a randomized trial with 92 primary schools, they find positive effects on language, after one year, but no statistically significant effects on math.

The remainder of this paper proceeds as follows. The next section briefly describes theoretical considerations. Section 3 gives a short introduction to Chile's teacher evaluation system, and Section 4 introduces the proposed estimation framework. This is followed by a description of the paper's data sources, in Section 5. Subsequently, Section 6 provides summary statistics for the analytic sample and scrutinizes the study's internal validity. Section 7 presents results and Section 8 concludes.

2. Theoretical Considerations

From a theoretical viewpoint, there are no clear expectations considering whether, if at all, teacher evaluations have a positive or negative impact on teacher effectiveness. A short review of five theoretical lenses illustrates this point. To begin, as discussed by Papay (2012) and Taylor and Tyler (2012), a *human resource development view* predicts increased teacher performance and improvements in student learning, as evaluations provide teachers with information on how to improve. This approach also suggests that evaluations allow teachers to learn about skill and performance expectations. Further, a *professionalization argument* hypothesizes that student learning may be improved if increases in teacher evaluations allow teaching to graduate from a "second grade" to a "full" profession (Johnson and Fiarman, 2012; Mehta, 2013).[7] Next, *principal agent theory* may also predict positive effects through improved information on effort, such as an increase in the ability of principals and parents to monitor effort and performance (Hölmstrom, 1979; Milgrom and Roberts, 1992). At the same time, *the multi-tasking model* (*cf.* Jacob, 2005) posits that, while evaluations may increase teachers' efforts to improve on those tasks that are observed by the performance measure, teachers may shift their efforts away from other, unobserved tasks. Thus, under this model, ambiguous effects can be expected. Finally, critics of teacher evaluations point to rather practical concerns and

[7] Mehta (2013) argues that teacher evaluations may also hinder professionalization if they are used to promote external teacher accountability based on test-scores; he promotes approaches instead.

to an *opportunity-cost argument*, suggesting that teacher evaluations may take up scarce financial resources and teachers' work-time (*cf.* Taut et al., 2011).

3. Teacher Evaluation in Chile

This section provides a short overview of Chile's teacher evaluation system, Docentemás, focusing on those characteristics that inform the study's identification strategy.[8] Docentemás was introduced in 2003 as a standards-based, formative assessment system that is tied to the country's national Framework of Good Teaching ("Marco para la Buena Enseñanza").[9] In 2005, participation became mandatory, for all public schools in the country.[10]

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%).[11] 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. The continuous score ranges from 1 to 4, and values of 2, 2.5, and 3 are used as cut-scores, respectively. However, a teacher's rating may be modified by a Municipal Evaluation Commission

[8] See a recent OECD review for a comprehensive presentation, including information on Chile's school system, in English (Santiago et al., 2013). See Manzi et al. (2011) for a detailed presentation of Chile's teacher evaluation system, in Spanish.

[9] The Framework is based on Danielson's Framework of Good Teaching and the Measures of Effective Teaching (MET) Project (see Santiago et al., 2013).

[10] For 2010, Manzi et al. (2011, 26) report that 96% of all public teachers complied with their legal obligation to participate in the evaluation. I calculate that, in 2015, 80% of eligible teachers had been evaluated at least once. This calculation focuses on elementary teachers in public schools, teaching either mathematics, reading, or "general".

[11] These weights change in the case of follow-up evaluations after a rating in the bottom category. The adjusted weights are as follows: Self-evaluation (5%); third-party reference report (5%); peer evaluator interview (10%); teacher performance portfolio (80%).

before it becomes final (modifications occur in approximately five percent of cases).

The overall evaluation spans one year. Its process begins with a teacher's nomination in April and continues with the submission of portfolios, recordings, self- and peer-evaluations between August and October, as well as with the third-party report in November. Grading takes place in December and January, final grades are decided upon in February and March, teachers receive their results in March with detailed written feedback, and further reports are distributed to other parties in April.[12] Interestingly, results for the largest evaluation component (centralized, anonymous ratings of the teacher performance portfolio) thus only become available *after* the remaining three components have been scored.[13]

In 2011, a new law (Ley 20.501) introduced changes to the consequences of a teacher's performance rating. Generally, public teachers are required to be evaluated at least once every four years.[14] Yet, under the new law, teachers rated as "basic" must be re-evaluated after two years.[15] The law came into effect in 2011, but it did not affect teachers retroactively, based on their previous performance ratings. Below (in Section 7.1), I show that the policy sharply increased a "basic" teacher's probability of getting re-evaluated after two years. In contrast, teachers in the pre-policy period and teachers with a higher rating

[12]The Chilean school year begins in March and ends in December.

[13]Arguably, this feature reduces the likelihood of score manipulation around the three cut-scores—I discuss this matter and its implications for the paper's econometric strategy further below.

[14]Docentemás covers all teachers in municipal schools above a set workload threshold. Teachers are nominated for evaluation by the head of their respective municipal school authorities ("Municipal Education Administration Department" or "Municipal Education Corporation"). New hires are not evaluated in their first year of service. Since 2006, teachers may opt out in their last three years before qualifying for retirement.

[15]Teachers with an "unsatisfactory" rating have to be re-evaluated directly in the following year and their contracts are terminated if their rating does not improve. This requirement did not change with the 2011 law. Yet, before 2011, "unsatisfactory" teachers were only dismissed if their rating did not improve in two subsequent evaluations, rather than one.

were not re-evaluated.

For "basic" teachers, the new law did not result in other changes—whether with respect to their job security, their access to incentive schemes, or their professional development, for example. In terms of teacher turn-over, since 2011, a teacher with a "basic" rating must leave the system if her rating does not improve in the next two assessments. However, the potential reduction in a "basic" teacher's job security only applies after her *second* follow-up evaluation. Since 2011, some principals are also allowed to dismiss up to five percent of "basic" *and* "unsatisfactory" teaching staff. Yet, this change only applies to a subset of principals who have been hired through a competitive process. Further below, I investigate—and do not find support for—the law's impact on "basic" teachers' turn-over. Teachers in the top two categories also receive access to a rewards and incentive scheme.[16] In contrast, "basic" teachers are barred from applying to this program. Further, teachers in the bottom two categories may be asked to participate in professional development activities before their next evaluation takes place.[17] Yet, crucially, assignment mechanisms for these programs did not change with the 2011 law.

Therefore, the new law affected teachers rated as "basic" (as opposed to "competent" or "outstanding") chiefly through the requirement to undergo a renewed evaluation two years later. My analyses are thus able to focus on a comparison of teachers whose per-

[16]Chile's evaluation framework consists of multiple components, which are chiefly the teacher performance evaluation system, Docentemás, the Program for the Variable Individual Performance Allowance (AVDI), the Program for the Accreditation of Pedagogical Excellence Allowance (AEP), and the National System for Performance Evaluation (SNED). AVDI represents a complementary, voluntary, reward system that is open to those municipal instructors rated within the top two of four performance brackets, as determined by Docentemás. AEP, on the other hand, provides an additional, voluntary reward system for all teachers, offering a monetary award to selected candidates, public praise, and the opportunity to apply to the "Maestros" Teacher Network. Lastly, SNED uses national test score data to offer group level incentives to schools (excluding private schools).

[17]These Professional Development Plans (PSPs) are paid for centrally, organized by municipalities, and mainly consist of courses, workshops, and seminars. See Cortés and Lagos (2011) for a detailed description of PSPs and related descriptive statistics, in Spanish. See Lombardi (2019) for an evaluation of their effectiveness.

formance score suggested a "basic" rating (inducing them to undergo a new evaluation) with a counterfactual situation in which they would have obtained a higher score, before and after the law was passed.

4. Identification Strategy

This study uses a fuzzy difference-in-difference ("fuzzy DD") estimation strategy. I exploit that (a) under the new law, teachers below the cutoff were induced to be re-evaluated (in contrast to teachers just above the cutoff), and (b) other effects of a "basic" (in contrast to a higher rating) rating stayed the same over the same period. The estimator moreover accounts for the fact that the law is not adhered to perfectly (in other words, the post-policy jump in the probability of getting evaluated after two years is "fuzzy").

More formally, I estimate a two-stage least squares (2SLS) regression, whose reduced form is given in Equation 1, as follows.[18]

$$Y_{j(t+x)i} = \beta_{RF0} + \beta_{RF1} T_{jt} + \beta_{RF2} TxPost_{jt} + \Gamma_t + \Omega + \boldsymbol{X}_{jti} + \epsilon_{j(t+x)i} \tag{1}$$

Reduced-form (RF) Equation 1 refers to teacher j, initially evaluated in year t. Here, Y denotes an outcome of interest (e.g., test scores) for teacher j's student i, x years past t. T is an indicator for being below the assignment breakpoint at any point of time, and $TxPost$ is an indicator for being below this breakpoint in the post-policy period (reflecting assignment to re-evaluation as per the continuous evaluation score teacher j received in year t). Γ_t captures year fixed effects; Ω captures commune fixed effects (I omit a commune subscript throughout); $\boldsymbol{X}$ is a vector of teacher and student character-

[18] This notation captures that a fuzzy estimation strategy is equivalent to an instrumental variable (IV) approach. My endogenous variable is teacher (re-)evaluation in year $t+2$, which is instrumented with an indicator for being below the breakpoint, in the post-policy period.

istics measured in baseline year t.[19] The respective first-stage (FS) equation (not shown) is equivalent to Equation 1, but now the outcome variable $Y_{j(t+2)}$ reflects a teacher's re-evaluation in year $t+2$, the estimation occurs at the teacher-level, there is hence one observation per teacher in year t, any subscripts i are therefore dropped, and the vector of baseline covariates $\boldsymbol{X}$ excludes student characteristics.

The coefficient of main interest is β_2. In the remainder of the paper, all reported effect sizes (and their standard errors) represent the Wald estimate β_{Wald2}, where $\beta_{Wald2} = \beta_{RF2}/\beta_{FS2}$. Given the (complete lack of) re-evaluations in the pre-period and near-zero re-evaluation rates for the post-period comparison groups (see Section 7.1 below), β_{Wald2} is interpreted as a Treatment-on-the-Treated (ToT) effect. In the paper's analysis of heterogeneous effects, I furthermore include interactions between a continuous measure of teacher experience and each of the three variables T, $Post$, and $TxPost$. In the following, β_6 refers to the coefficient on the interaction between $TxPost$ and teacher experience.[20]

I calculate the study's two-stage least-squares Wald estimates through a bootstrap procedure (with 750 replications) and obtain clustered standard errors by blocking resamples at the teacher-year level. I repeat this procedure for each outcome variable and for the three points in time after a teacher's assignment (that is, $x = 1$, $x = 2$, and $x = 3$).

A final comment is in order as the availability of a continuous assignment variable with a cut-off rule may have—misleadingly—pointed to a simple regression-discontinuity (RD) strategy. However, recall that a regression-discontinuity strategy would not account for the fact that other Chilean programs use the same cut-off to determine eligibility. Additionally, a simple RD approach assumes that teachers' assignment to treatment is as

[19] Following common approaches to model student growth trajectories, all student controls also include the quadratic of a child's baseline GPA (cf. Singer and Willett, 2003).

[20] The respective Wald estimate is calculated as follows: $\beta_{Wald6} = (\beta_{RF2} + \beta_{RF6})/(\beta_{FS2} + \beta_{FS6}) - \beta_{Wald2}$.

good as random at the threshold score, which implies that teachers are not able to manipulate their scores around this cut-off. Finally, an earlier version of this paper pursued a fuzzy difference-in-discontinuities (or fuzzy difference-in-RD estimator)—it was abandoned, due to lack of statistical power.

5. Data

For the years 2005 to 2015, I use teacher-classroom links to match administrative data for the universe of elementary teachers in Chile's public schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students, and results on standardized test scores (in mathematics and language) for all Chilean fourth graders attending public schools. More precisely, I combine data from five different sources.[21] The first data source is the "Ideoneidad Docente" data-base, which is maintained by Chile's Ministry of Education. The data-base includes detailed, administrative information on the population of Chilean teachers such as 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), information on a teacher's training (such as subject specialization and the training institution), identifiers for the school and grade level a teacher taught in a given year, and information on the school (such as whether a school is located in an urban or in a rural area).

Second, the above data is merged with a data-set containing information on whether a teacher participated in Docentemás in any given year between 2005 and 2015. This data-set also includes detailed information on each teacher's final performance rating, the continuous performance score, and her rating on each of the four evaluation compo-

[21] If not indicated otherwise, data-sources are in the public domain and can be downloaded from a website maintained by the Education Ministry's "Centro de Estudios" (2016). Data-sets are merged by using unique school identifiers, information on grade levels and classes, and unique (codified) teacher identifiers.

nents. The study's third data-source consists of the "Asignatura por Docente" data-set for 2005-2015, which provides information on the class(es) and subject(s) a teacher taught in a given year.[22] Fourth, the study combines administrative information on the universe of Chilean students, their absentee rate (as a percentage of school days), whether they repeated a given school year, and their end-of-year grade point average (GPA).

Fifth, student learning outcomes are measured using Simce exam scores.[23] 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, standardized exam administered at the end of the school year (across private, subsidized, and public schools). As of 2015, Simce has covered a wide array of subjects and levels, but most notably fourth-grade mathematics and (Spanish) language in every consecutive year since 2005. Simce data-sets 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). Given the salience of Simce scores in Chile, I do not transform them to standard deviations. However, results remain easily interpretable as Simce test-scores are scaled to a standard deviation of 50.

For each year, Simce assessments are also complemented by a student, a parent, and a teacher survey. For a subset of years, these surveys provide comparable measures of potential mediating factors: teaching effort, teachers' level of caring, and teacher beliefs. For the years 2012 through 2015, I construct an index of student-reported teaching practices, or "effort". More specifically, I calculate an average over six survey questions that

[22]This data-set is not in the public domain. The Ministry asks researchers to undergo a standardized process to receive access to the data-set.

[23]The student-level version of this data-set is not in the public domain. The Ministry asks researchers to undergo a standardized process to receive access to the data-set.

seek to capture a teacher's classroom behaviors.[24] Moreover, for 2011 to 2014, I use a measure that asks parents to rate the extent to which their child's head teacher cares about her students.[25] As there is no head teacher identifier, in my analyses of mechanisms, I assume that in fourth grade, a teacher is considered the head teacher if she teaches both math and language.[26] Finally, for each year from 2005 to 2014, surveys ask teachers to state their beliefs concerning students' future educational attainment.[27]

6. Sample Characteristics and Internal Validity

6.1. Sample

Table 1 provides summary statistics for the study's sample of teachers at "baseline" (that is, the year of their initial evaluation, t), and the analysis sample of teachers observed teaching grade four students two years later (in year $t+2$). This includes teachers who were initially evaluated between 2005 and 2013, and potentially re-evaluated between 2007 and 2015. Appendix Table A1 presents the respective descriptives for years $t+1$ and $t+3$.[28]

[24]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. Results (available upon request) are robust to using an alternative index from a principal component analysis instead (with a polychoric correlation matrix, extracting the first joint component from the six items).

[25]Rated from 1 or "very unsatisfied" to 7 "very satisfied".

[26]Approximately, 90 percent of the study's sample of fourth graders have the same math and language teacher. I 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, I average the student responses across subjects.

[27]Teachers choose one of the following six categories: 1) Will not complete eighth grade, 2) will complete eighth grade on the technical-professional track, 3) will complete eighth grade on the humanist-scientific track, 4) will complete a technical degree, 5) will complete a university degree, 6) will complete postgraduate studies.

[28]In 2016, Chile introduced major changes to teachers' career pathway, including in the way teacher evaluations are used (*Ley 20.903*). This suggests that data for 2016 and thereafter should not be used for the present analysis. At the same time, I do not have access to data for these years. Teachers who were initially evaluated in 2013 (and their students) are thus not observed in $t+3$.

I restrict all analyses to teachers who were initially evaluated in elementary and I drop those teachers who would have been too old to be eligible for re-evaluation two years later.[29] I also drop the small share of approximately 1.8 percent of teachers with a performance score that would have suggested an "unsatisfactory" rating.[30] This approach renders 31,400 teacher-by-year observations (22,435 for the pre-policy period and 8,965 for the post-policy period). Of those, 7,458 represent teachers of grade four students in language or mathematics, two years after the teacher's assignment to re-evaluation.[31] This mapping of teachers to their students results in 157,153 year-teacher-student observations.[32]

[Table 1 about here]

6.2. Threats to Internal Validity

In the following, I investigate—and present evidence against—five potential threats to the study's internal validity: differential attrition based on a teacher's assignment status (as per her initial evaluation score in year t); imbalance of observable teacher characteristics across teachers "assigned to treatment" and their comparison group; potential sorting of students to (or away from) teachers who are assigned to be re-evaluated; whether, with the policy change, teachers manipulated their assignment status more (/less); and whether the two groups of teachers (those to be "assigned to treatment" and their comparison group) exhibited differential pre-trends.

[29] Teachers within three years of the retirement age are not required to be evaluated.

[30] It is unclear whether, due to the new law, informal re-evaluation criteria may have changed for these teachers. I do not use other criteria to drop teachers (such as the teacher's workload), as these may have changed post-assignment.

[31] These 31,400 teacher-by-year observations comprise 22,066 unique teachers, of whom 6,844 teach fourth-grade language or mathematics.

[32] These year-teacher-student observations comprise 119,382 unique students. Students may be observed twice, either if math and language are taught by different teachers, in a given year, or if students repeat fourth grade.

6.2.1. Differential attrition

Table 1's last column follows the paper's difference-in-difference strategy (as described in Section 4 above, with the exclusion of covariates). The Table's top panel reports whether the introduction of the law coincided with systematic changes in "attrition" rates—i.e., whether students became systematically less (/more) likely to be taught by a "basic" mathematics or language teacher. Table 1's findings suggest that, in the post-policy period, "basic" teachers became slightly less likely to teach a fourth-grade class in mathematics or language in the year after evaluation scores are released (by two percentage points, statistically significant at the 0.1 level). If these weaker teachers are thus systematically removed, Equation 1 may be slightly under-estimating the true effect of teachers' re-evaluations, for year $t+1$. For the year of the re-evaluation ($t+2$) and the year thereafter ($t+3$), in contrast, I do not find evidence of systematic removal (or assignment) of "basic" teachers to fourth-grade math and language classrooms, as the new law was introduced.

6.2.2. Baseline balance for teachers

For fourth-grade Simce teachers, there are only negligible differences in teachers' gender or age, their contract hours, their years of experience, and in the percentage of teachers who work in more than one school. At baseline, I also find no differences in teachers' average school-level Simce scores (whether in fourth-grade math or language). As an exception, three years post assignment, assigned teachers were slightly less experienced, by 2.4 years (significant at the 0.05 level; see Appendix Table A1). This difference is not confirmed for the remaining two samples. This finding therefore appears to reflect multiple hypothesis testing, rather than systematic differences. Yet, I also control for these teacher characteristics, including a teacher's years of experience, in the vector of baseline covariates.

6.2.3. Baseline balance for students

To investigate the potential of systematic sorting of students, Table 1 also includes descriptive statistics for teachers' fourth-grade Simce-taking students, two years post-assignment (yet measured at baseline, in year t).[33] The table also presents additional information on student demographics (household income and the mother's highest level of education), even though this information is not available for all students, it is measured at the time of follow-up, and in all of the paper's regression analyses, these variables are therefore not included as covariates.

I find no support for the hypothesis that schools engage in sorting of students to (or away from) teachers who are assigned to be evaluated. None of the tests point to differences in student characteristics (at the 0.1 level). Moreover, point estimates are close to zero, with tight error bands, suggesting that students' prior academic achievement, grade retention, attendance, gender, household income, and maternal level of education are balanced as the difference among groups below and above the assignment threshold is compared across the pre- and post-policy periods.

6.2.4. Differential manipulation of assignment status

Another concern revolves around whether, with the policy change, teachers close to the threshold for a "basic" rating were systematically assigned to more (/less) lenient performance ratings. As teachers, their peers, and principals determine 40 percent of the performance score (through self- and peer-evaluations, as well as reference-reports), there may have been an increase (/decrease) in the share of teachers whose score—and thus treatment assignment—was manipulated. Two facts alleviate this concern.

[33] Appendix Table A1 provides the respective information for the samples one year and three years past baseline.

First, recall that the remaining 60 percent of a teachers' continuous score is based on her portfolio, which is rated centrally, anonymously, and *after* the remaining three components have been scored. For a teacher, her peers or the principal, it is thus impossible to know whether the composite score will be close to the cut-off.[34]

Secondly, in Figure 1, I build on work by McCrary (2008) to assess empirically whether score manipulation around the cut-off score differed across the pre- and post-policy periods.

[Figure 1 about here]

The figure's top-panel shows McCrary plots for the pre-period (left) and the post-period (right). These plots are generated by calculating a finely-gridded histogram, which is then smoothed using local linear regression, separately on either side of the breakpoint. A formal test (McCrary, 2008) on the difference-in-densities around the breakpoint does not reject the null of equal densities on both sides, for either period (at the 0.01 level). Further, this paper's identification strategy simply posits that, if present at all, the extent of manipulation remained unaffected by the policy change. In the bottom panel, I extend McCrary's (*ibid.*) method by calculating the difference-in-difference of densities, for common bin-sizes of 0.01 points[35], and smoothing over the histogram thereafter (separately, for both sides of the breakpoint).[36] The bottom panel illustrates how the difference in densities around the breakpoint remains constant over time (not significantly different from zero at the 0.1 level). In summary, this graph (and

[34] See Lombardi (2019), for a similar argument.

[35] Teacher evaluation scores are reported in increments of 0.01 points.

[36] To my best knowledge, this is the first study presenting a McCrary plot for the difference-in-differences of densities. However, I do not calculate optimal bin sizes and deviate from McCrary's (2008) method of choosing the optimal bandwidth. I choose a bandwidth of 0.2 points and a bin size of 0.01 points, as in the remainder of the paper. For consistency with McCrary's (*ibid.*) method, I include a fourth-order polynomial on both sides of the breakpoint. I thank Ugo Troiano for helpful comments.

the respective test of a difference-in-difference of densities) thus shows that a difference-in-differences approach alleviates concerns regarding manipulation around the cut-off.

6.2.5. Common trends

As with any difference-in-difference estimation, the identifying assumption is that the average change in the comparison group's outcomes represents the counterfactual change in the treatment group's outcomes (in the absence of treatment). While not directly testable, I present pre-policy trends in outcome variables, for teachers assigned to a basic rating vs. teachers assigned to a higher rating (in year $t+2$). Appendix Figure A1 shows that, in the pre-policy period, the explained (left panel) and unexplained (right panel) portions of test score variance follow parallel trends, across these two groups of teachers. I therefore conclude that there is no evidence to suggest a violation of the common trends assumption.

7. Results

7.1. First Stage Results

Figure 2 provides evidence for the validity of the study's first stage. Each point represents the share of teachers being re-evaluated after two years, in score bins with a width of 0.02 score points. The solid line plots predicted values, with separate linear trends estimated on either side of the basic vs. competent (or better) cut-off. This threshold is indicated by the red, vertical line. The dashed lines show 95 percent confidence intervals. In the pre-period (left panel), the percentage of teachers who are newly evaluated in year $t+2$ is consistently zero. Thus, by including the pre-period, the proposed estimator solely differences out potential effects that occurred around the same threshold (for example, through eligibility for incentives or training, rather than an effect of re-evaluations).

In the post-period (right panel), as expected, a large jump in the probability of re-evaluation occurs around the breakpoint. Teachers' predicted share of re-evaluation just to the left of the threshold (suggesting a "basic" rating) is 63 percent; in contrast, the percentage remains close to zero once the threshold is crossed (0.2 percent). Note that there is great variance with respect to compliance (or the level of "fuzziness") among the assigned teachers. Yet, in addition to this visual evidence, the formal estimate of Equation 1 also confirms the strength of the first stage relationship—the F statistic for a test of $\beta_{FS2} = 0$ is 3843.4, for the sample of teachers teaching fourth-grade two years after the assignment. This formal estimate suggests that the law increased the probability of re-evaluation by 62.8 percentage points.[37]

[Figure 2 about here]

7.2. Effects on Student Learning

Table 2 shows the study's main results, for the investigation of effects on student learning. In Table 2, coefficients for β_1, year and commune fixed effects, and the vector of teacher and student characteristics are omitted. Models in odd-numbered columns do not account for potentially heterogeneous effects by teacher's level of work-experience. Even-numbered columns refer to models that interact the treatment with a teacher's years of work-experience. Recall that β_{Wald2} captures the main ToT effect of a teacher's re-evaluation, whereas β_{Wald6} reflects the additional ToT effect, times the teacher's years of experience in year t.[38] Recall also that, in each year and subject, Simce scores are scaled to a standard deviation of 50.

[Table 2 about here]

As shown in Column (5), I find that students who were taught by a teacher who was re-evaluated one year prior did not perform differently, compared to their peers,

[37] Results for those teachers observed teaching fourth-grade in $t+1$ and $t+3$ are similar and available upon request.

[38] Given the two-stage least-squares set-up, I do not report R^2.

whether in math or reading. Column (3) also does not lend support for the hypothesis that there is a detrimental effect of teacher evaluations on student test scores. Column (1) reports findings for the year prior to a teacher's evaluation, one year after the "treatment" was assigned ($t+1$). In this year, teachers may have changed their behavior once they learned about their treatment status (Ashenfelter, 1978). Yet, for both subjects, I do not find such an effect.

Columns (2), (4), and (6) assess whether these results differ for teachers with fewer (or more) years of work experience. The results do not support such a phenomenon; the coefficients of β_{Wald6} are statistically indistinguishable from zero. Yet, for language and year $t+1$, I find a negative effect among new teachers (approximately 0.17 standard deviations in expectation, significant at the 0.1 level). This effect is of similar size for mathematics but not statistically significant. In summary, I therefore conclude that teacher's re-evaluations did not lead to increased teacher productivity (as measured by student test scores). I find that this observation is independent from a teacher's level of work experience.

7.3. Effects on Teachers and Teaching

Table 3 reports on effects on teachers' (student-reported) teaching behaviors, (parent-reported) levels of caring, and (self-reported) beliefs in their students' future educational attainment. All measures are standardized. All models follow Equation 1; they include a vector of baseline teacher and student characteristics, commune and year fixed effects, and cluster standard errors at the teacher-year level.

[Table 3 about here]

Table 3 suggests negative effects on teaching practices in the year of a teacher's re-evaluation, and in the year thereafter (of 0.22 and 0.30 standard deviations, respectively; significant at the 0.01 level). The table also documents negative effects on teachers' levels

of caring, in the year after the teacher's assignment (of 0.24 standard deviations; significant at the 0.05 level). I do not find other effects for the remaining year-outcome combinations. Moreover, none of these results differ by teachers' level of work experience.

7.4. Robustness of Findings

I present three types of robustness checks. I present evidence for the absence of effects in the pre-period (i.e., $t-1$, the year prior to a teacher's potential assignment). I moreover re-estimate the study's main results by adding a group-specific, linear time trend for "basic" teachers to Equation 1. Finally, I re-estimate the study's main results by slightly modifying the assignment indicator. Results from these checks are presented in Table 4 below.

[Table 4 about here]

7.4.1. Falsification test

Column (1) of Table 4 presents results from a falsification test. In this test, I estimate Equation 1 for students one year *before* their teacher was subjected to her initial evaluation (and thus potentially assigned to be re-evaluated). I do not find evidence for "effects" of a teacher's later assignment. This result further corroborates the findings from Section 6.2.5, above, and alleviates concerns of differential pre-trends.

7.4.2. Robustness under inclusion of time trends

Columns (2) to (4) of Table 4 present a re-estimation of Equation 1 with the inclusion of a group-specific linear trend, for teachers with an evaluation score that suggests a "basic" rating. This check is available for learning outcomes and teachers' beliefs only; the remaining two outcomes only provide data for two (/three) pre-policy years, respectively. More specifically, I maintain Equation 1's year fixed effects and add an interaction of T with a continuous year indicator.

The results from this analysis confirm the findings reported in Section 7.2 (see Table 2, odd columns) for math, and for language in the year of and the year after a teacher's re-evaluation. In the year prior to the re-evaluation, however, we now observe negative effects on language (of .27 standard deviations). All remaining effects on instruction, teachers' level of caring, and beliefs are indistinguishable from zero.

7.4.3. Robustness to alternative treatment assignment

So far, the article's analyses have used the continuous score to reconstruct whether a teacher was to be assigned for re-evaluation. Columns (5) to (7) of Table 4 present findings from a re-estimation of Equation 1 if a teacher's actual performance rating is used for this purpose instead.[39] These results document a negative effect (of 0.11 standard deviations) on language learning in year $t+1$, the year after assignment and prior to re-evaluation. The respective effect for math is of similar size but statistically insignificant. For that year, they also suggest negative effects (of 0.16 standard deviations) on teachers' levels of caring. Finally, the results point to negative impacts on teaching practices both in the year of and in the year after a teacher's re-evaluation (of 0.16 and 0.27 standard deviations, respectively). All remaining coefficients are indistinguishable from zero.

7.5. Bounding of positive ITT effects on student learning

Are the above null-findings on student learning precise enough to rule out positive effects of Chile's policy, which mandates repeat teacher evaluations? To investigate this question, I now switch the study's focus to intent-to-treat (ITT) estimates. I repeat the above bootstrap procedure, plot the distribution of ITT coefficients (from 750 draws), and report on the 95th percentile. I interpret this value as the upper bound of ITT effects, and rule out higher impacts. To gain further precision, in doing so, I follow de Ree et al. (2018) and pool observations across subjects.

[39] Recall that a local commission may object to a teacher's score and override her rating. Note how this alternative assignment definition may thus not be entirely exogenous.

Figure 3 reports on the results from this bounding exercise. The top panels reports on the ITT effect in the year of the evaluation (left panel) and in the year after the evaluation (right panel). The bottom panels allow for heterogeneity in impacts and report on the corresponding ITT effects for teachers with just one year of work experience.

[Figure 3 about here]

Positive ITT effects are ruled out precisely. With 95 percent confidence, the results reject effects larger than 0.03 standard deviations for the year of the evaluation, and of 0.08 standard deviations for the year after the evaluations. For teachers who just started teaching, and may be expected to be most receptive of personal development, these bounds are even smaller (0.02 standard deviations and 0.06 standard deviations, respectively).

8. Conclusion

This study offers quasi-experimental evidence on the effects of formative performance evaluations on teacher effectiveness and child learning. In summary, I cannot conclude that Chile's repeat performance evaluations lead to substantial gains in student achievement, one year after a teacher is assigned to be evaluated. For the year of the evaluation, I also do not find effects on student learning. Positive impacts are ruled out precisely. Instead, the study results suggest that concerns about detrimental effects may be at least partly warranted. Some specifications point to negative effects on language learning, in the year after a teacher's assignment. The paper moreover documents decreases in teachers' level of caring, for the same year. I also observe additional negative effects on teaching practices in the year of and in the year after a teacher's evaluation. I do not detect impacts on teachers' beliefs in their students' future educational attainment.

This study isolates the effect of *repeat* evaluations; it cannot speak to the effects of a teacher's initial evaluation. Yet, for policy makers considering the introduction of a national system, this is arguably the more important question to consider: Will teachers' performance and student learning improve as teachers are regularly subjected to evaluations?[40] Taken together, in evaluating the impact of repeat evaluations, this study aims to provide first evidence on this question—for a comprehensive, standards-based teacher evaluation system that has been described as a role model for other countries (Bruns and Luque, 2014). To my best knowledge, it is only the second rigorous study on the effects of formative performance evaluations, and the first analysis under a well-established evaluation system that operates at national scale.

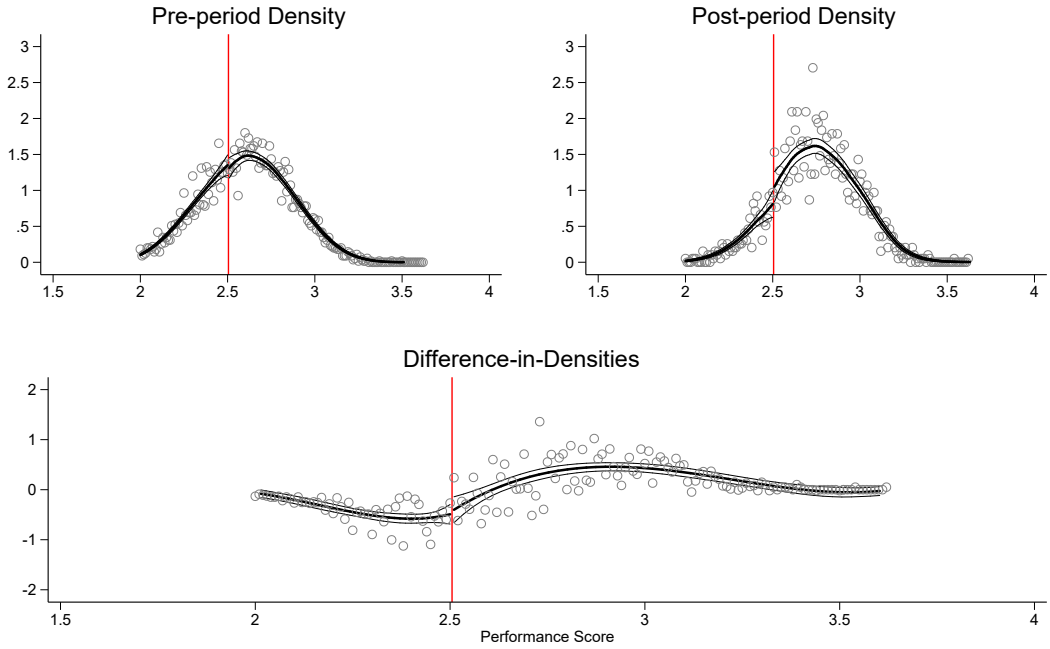
As discussed by Taut et al. (2011), Chilean policy makers regularly re-consider whether "Docentemás" is worth its cost and whether the system should be expanded to private schools (Educación 2020, 2013). Moreover, given the scale and nature of the investigated program, even decision makers in other public sectors may look to the example of Docentemás as staff performance evaluation systems are (re-)considered. In the light of these debates, this article casts doubt on the use of repeated formative evaluations as a means to improve employee productivity.

[40] Compare to Taylor and Tyler (2012, 3629), who also stress this point.

Figures and Tables

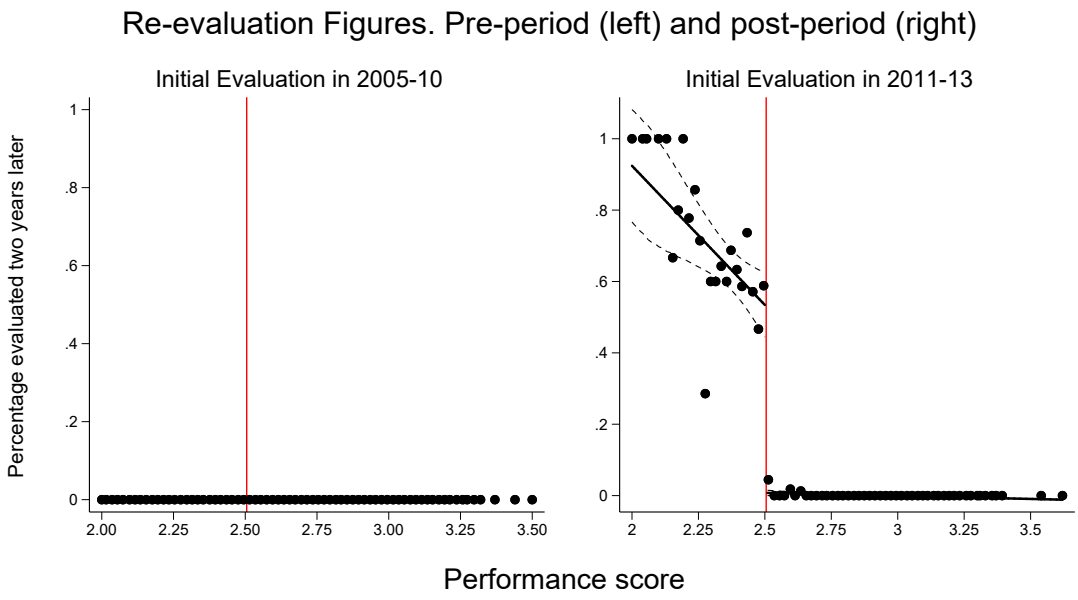
Figures

Figure 1: McCrary Plots



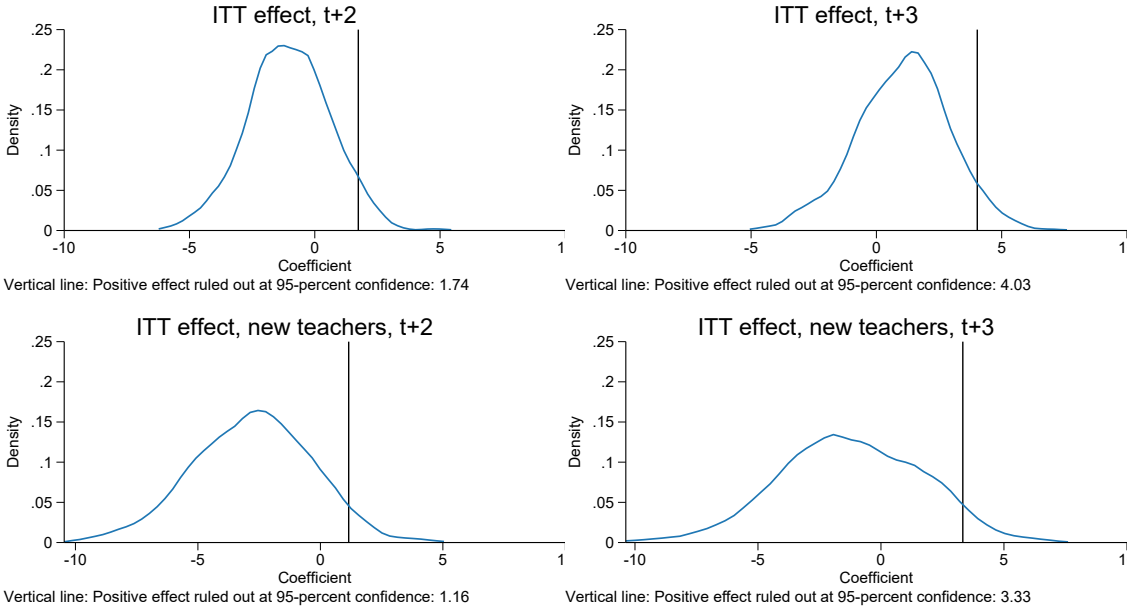
Note: Vertical lines indicate the recommended cutoff score.

Figure 2: First Stage



Notes: Each point represents the share of teachers being re-evaluated in score bins of width 0.02 score points. The solid line plots predicted values, with separate linear trends estimated on either side of the basic vs. competent threshold. This threshold is indicated by the vertical line. The dashed lines show 95 percent confidence intervals. No predicted values or confidence intervals shown for the pre-period as the share is consistently zero.

Figure 3: Bounding of positive ITT effects on student learning



Notes: This figure provides kernel density plots of ITT coefficients, from a bootstrap with 750 draws, clustered at the teacher-year level. Vertical lines indicate the 95th percentile. 'New teachers' refers to teachers with one year of experience. Regressions include year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).

Tables

Table 1: Sample Characteristics and Validity Checks

	2005-2010		2011-13		
	Below	Above	Below	Above	DD
Attrition					
In sample in t+1	0.18	0.21	0.15	0.21	-0.02 (0.01)*
In sample in t+2	0.17	0.21	0.15	0.21	-0.02 (0.01)
In sample in t+3	0.16	0.2	0.11	0.1	0.01 (0.01)
n	8,192	14,243	1,599	7,366	31,400
Teacher Baseline Characteristics					
Gender: Female	0.75	0.83	0.77	0.86	-0.03 (0.03)
Contract hours	38.56	38.28	38.11	37.35	0.67 (0.50)
Works in yet another school	0.07	0.06	0.04	0.02	0.01 (0.02)
Years in service	19.76	18.66	14.23	14.96	-0.83 (0.79)
n	1,886	3,612	279	1,681	7,458
School's baseline reading score†	241.68	247.83	249.85	254.95	-0.35 (1.53)
School's baseline math score†	230.4	236.01	238.96	246.52	-1.74 (1.72)
n	1,387	2,928	235	1,533	6,083
Student Baseline Characteristics					
GPA	5.89	5.95	5.87	5.91	0.03 (0.02)
Repeated in baseline year	0.03	0.03	0.04	0.04	0.00 (0.00)
Attendance	92.64	92.97	90.78	91.65	-0.24 (0.25)
Gender: Female	0.49	0.49	0.48	0.49	-0.00 (0.01)
n	36,410	79,242	5,391	36,110	157,153
Household income (pesos)††	265977	278862	343669	328037	-1242.60 (12244.54)
Mother's edu. (years)††	9.85	10.18	10.76	10.81	-0.08 (0.11)
n	29,637	66,198	4,655	31,666	132,156

Notes: "Teachers" include all unique year-teacher observations and may thus repeatedly include individual teachers over time. "Students" include all unique year-teacher-student observations and may thus include up to two observations per student and year (if math and reading are taught by different teachers, in a given year). Teacher and student baseline characteristics refer to the analysis sample observed at $t+2$. Appendix Table A1 reports on the sample observed for $t+1$ and $t+3$. "Below" and "Above" refer to teachers below or above the cut-off, respectively. t refers to the year of the initial evaluation. All variables measured in t, if not denoted otherwise. † denotes variables available for fewer observations (and not included as covariates). †† denotes variables measured at follow-up (and not included as covariates). Note that the 2013 sample is not followed up in $t+3$. "DD" refers to a difference-in-difference estimate as described in Section 4 (excluding control variables but including commune-level fixed effects). Standard errors in parentheses. For student-level characteristics, standard errors are clustered at the year-teacher level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: ToT Effects on Student Learning

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
β_{Wald2}	-3.045 (4.189)	-6.687 (5.668)	-2.040 (2.991)	-5.167 (4.730)	0.484 (3.887)	-1.274 (6.173)
β_{Wald6}		0.117 (0.264)		0.210 (0.256)		0.032 (0.301)
n (teachers)	6991	6911	6643	6565	5417	5336
n (students)	142515	142515	133013	133013	110000	110000
Language						
β_{Wald2}	-4.328 (3.591)	-8.522 (4.823)*	-1.966 (2.727)	-6.194 (4.224)	2.287 (3.486)	-4.232 (5.311)
β_{Wald6}		0.133 (0.224)		0.337 (0.260)		0.418 (0.274)
n (teachers)	7158	7076	6869	6785	5645	5556
n (students)	144868	144868	136938	136938	112986	112986

Notes: In odd columns, β_{Wald2} captures the ToT effect of a teacher's re-evaluation.
In even columns, β_{Wald6} captures the interaction (ToT) effect between a teacher's re-evaluation and her work experience.
Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).
Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: ToT Effects on Teaching Behaviors, Caring, Teacher Beliefs

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Practices						
β_{Wald2}	-0.110 (0.110)	-0.113 (0.164)	-0.220 (0.067)***	-0.261 (0.117)**	-0.295 (0.088)***	-0.215 (0.130)*
β_{Wald6}		0.000 (0.007)		0.004 (0.006)		-0.007 (0.008)
n (teachers)	2291	2267	3183	3142	2511	2477
n (students)	45207	45207	61689	61689	50315	50315
Caring						
β_{Wald2}	-0.242 (0.112)**	-0.276 (0.169)	0.050 (0.091)	0.059 (0.138)	0.006 (0.086)	0.072 (0.160)
β_{Wald6}		0.001 (0.008)		-0.000 (0.006)		-0.004 (0.009)
n (teachers)	2065	2044	2143	2117	1912	1889
n (parents)	40835	40835	41725	41725	39154	39154
Beliefs						
β_{Wald2}	0.106 (0.177)	-0.165 (0.247)	0.109 (0.164)	0.138 (0.234)	-0.085 (0.286)	0.076 (0.478)
β_{Wald6}		0.016 (0.011)		-0.002 (0.010)		-0.010 (0.029)
n (teachers)	5853	5788	4801	4749	4028	3972

Notes: In odd columns, β_{Wald2} captures the ToT effect of a teacher's re-evaluation.
In even columns, β_{Wald6} captures the interaction (ToT) effect between a teacher's re-evaluation and her work experience.
Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).
Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Robustness Checks

	Falsification	Group-specific time trends			Re-defining assignment		
	t-1	t+1	t+2	t+3	t+1	t+2	t+3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math							
β_{Wald2}	-0.982 (4.182)	-8.698 (5.871)	2.593 (4.648)	-2.567 (4.932)	-4.890 (3.342)	-1.480 (2.724)	-2.090 (3.305)
n (teachers)	4313	6083	5726	4627	6083	5726	4627
n (students)	100049	142515	133013	110000	142515	133013	110000
Language							
β_{Wald2}	1.748 (3.510)	-13.868 (5.243)***	3.027 (4.341)	-0.009 (4.419)	-5.784 (3.119)*	-0.752 (2.539)	-0.524 (2.859)
n (teachers)	4410	6218	5857	4853	6218	5857	4853
n (students)	101649	144868	136938	112986	144868	136938	112986
Practices							
β_{Wald2}	-0.016 (0.109)				0.027 (0.096)	-0.161 (0.065)**	-0.268 (0.069)***
n (teachers)	1027				2296	2985	2385
n (students)	19001				42016	55654	44025
Caring							
β_{Wald2}	-0.147 (0.127)				-0.157 (0.095)*	0.021 (0.082)	0.015 (0.075)
n (teachers)	955				2044	2031	1814
n (parents)	16946				37566	37840	34188
Beliefs							
β_{Wald2}	0.011 (0.161)	0.123 (0.197)	-0.051 (0.186)	-0.115 (0.223)	-0.012 (0.120)	-0.024 (0.132)	0.011 (0.180)
n (teachers)	4653	6536	5422	4603	6536	5422	4603

Notes: β_{Wald2} captures the ToT effect of a teacher's re-evaluation.
Column (1) reports on effects the year prior to assignment. Columns (2) to (4) add a time-trend for teachers with a "basic" evaluation score to Equation 1.
Group-specific time trends omitted for outcomes with data for only three years of pre-policy data, or less.
Columns (5) to (7) base a teacher's assignment on her final evaluation rating, not on the underlying evaluation score.
Not reported: Main effect, year fixed effects and a vector of baseline teacher and student characteristics (teacher's gender, age, contract hours, employment in another school, years of service, baseline school-level average Simce scores in math and reading; students' GPA and its square, attendance, retention).
Bootstrapped standard errors in parentheses (750 draws), clustered at the teacher-year level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Allen, J. P., R. C. Pianta, A. Gregory, A. Y. Mikami, and J. Lun (2011, August). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science 333*(6045), 1034–1037.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics 60*(1), 47–57.

Aucejo, E. M., T. F. Romano, and E. S. Taylor (2019). Does Evaluation Distort Teacher Effort and Decisions? Quasi-experimental Evidence from a Policy of Retesting Students. Working Paper 1612, Centre for Economic Performance, LSE, London.

Bergman, P. and M. J. Hill (2018, October). The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers. *Economics of Education Review 66*, 104–113.

Bruns, B. and J. Luque (2014, January). Great teachers: how to raise student learning in Latin America and the Caribbean. Technical Report 89514, The World Bank.

Burgess, S., S. Rawal, and Taylor, Eric S. (2019). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. Working Paper 19-139, Annenberg Institute, Providence, RI.

Centro de Estudios (2016, April). Base de Datos.

Cortés, F. and M. J. Lagos (2011). Consecuencias de la Evaluación Docente. In J. Manzi, R. González, and Y. Sun (Eds.), *La evaluación docente en Chile*, pp. 137–156. Santiago de Chile: MIDE UC, Centro de Medición Pontificia Universidad Católica de Chile.

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

Educación 2020 (2013, March). Opinión de Educación 2020 sobre la Evaluación Docente 2012.

Garet, M. S., A. J. Wayne, S. Brown, J. Rickles, M. Song, and D. Manzeske (2017, December). The Impact of Providing Performance Feedback to Teachers and Principals. Report NCEE 2018-4001, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Gersten, R., J. Dimino, M. Jayanthi, J. S. Kim, and L. E. Santoro (2010). Teacher Study Group Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms. *American Educational Research Journal 47*(3), 694–739.

Grissom, J. A. and P. Youngs (Eds.) (2016). *Improving teacher evaluation systems: making the most of multiple measures*. New York, NY: Teachers College Press.

Hölmstrom, B. (1979, April). Moral Hazard and Observability. *The Bell Journal of Economics 10*(1), 74–91.

Jackson, C. and J. Cowan (2018, December). Assessing the Evidence on Teacher Evaluation Reforms. Policy Brief 13-1218-1, National Center for Analysis of Longitudinal Data in Education Research, Washington, D.C.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5), 761–796.

Johnson, S. M. and S. E. Fiarman (2012). The potential of peer review. *Educational Leadership 70*(3), 20–25.

Kane, T. J., D. Blazar, H. Gehlbach, M. Greenberg, D. Quinn, and D. Thal (2019, April). Can Video Technology Improve Teacher Evaluations? An Experimental Study. *Education Finance and Policy*, 1–55.

Koedel, C., J. Li, M. G. Springer, and L. Tan (2017, April). The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. *American Educational Research Journal 54*(2), 241–278.

Koedel, C., J. Li, M. G. Springer, and L. Tan (2019, January). Teacher Performance Ratings and Professional Improvement. *Journal of Research on Educational Effectiveness 12*(1), 90–115.

Kraft, M. and H. Hill (2019). Developing Ambitious Mathematics Instruction Through Web-Based Coaching: A Randomized Field Trial. Working Paper 19-119, Annenberg Institute, Providence, RI.

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

Kraft, M. A., J. P. Papay, and O. L. Chi (2020). Teacher Skill Development: Evidence from Performance Ratings by Principals. *Journal of Policy Analysis and Management* (Forthcoming).

Lombardi, M. (2019, December). Is the remedy worse than the disease? The impact of teacher remediation on teacher and student performance in Chile. *Economics of Education Review 73*, 101928.

Louis, K. S. and H. M. Marks (1998). Does professional community affect the classroom? Teachers' work and student experiences in restructuring schools. *American Journal of Education*, 532–575.

Lovison, V. and E. S. Taylor (2018, September). Can Teacher Evaluation Programs Improve Teaching? Technical Report, Stanford University, Stanford, CA.

Manzi, J., R. González, and Y. Sun (2011). *La evaluación docente en Chile*. Santiago de Chile: MIDE UC, Centro de Medición Pontificia Universidad Católica de Chile.

McCrary, J. (2008, February). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

Mehta, J. (2013). *The allure of order: High hopes, dashed expectations, and the troubled quest to remake American schooling*. New York: Oxford University Press.

Milgrom, P. R. and J. Roberts (1992). *Economics, organization, and management*. New York: Prentice-Hall.

National Education Association (2019). Teacher Assessment and Evaluation: The National Education Association's Framework for Transforming Education Systems to Support Effective Teaching and Improve Student Learning. White Paper, National Education Association, Washington, D.C.

Papay, J. (2012, April). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review 82*(1), 123–141.

Papay, J. P., E. S. Taylor, J. H. Tyler, and M. Laski (2019). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*.

Pope, N. G. (2019, April). The effect of teacher ratings on teacher performance. *Journal of Public Economics 172*, 84–110.

Santiago, P., F. Benavides, C. Danielson, L. Goe, and D. Nusche (2013, November). *Teacher Evaluation in Chile*. Paris: Organisation for Economic Co-operation and Development.

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

Singer, J. D. and J. B. Willett (2003, March). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press.

Steinberg, M. P. and L. Sartain (2015, August). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy 10*(4), 535–572.

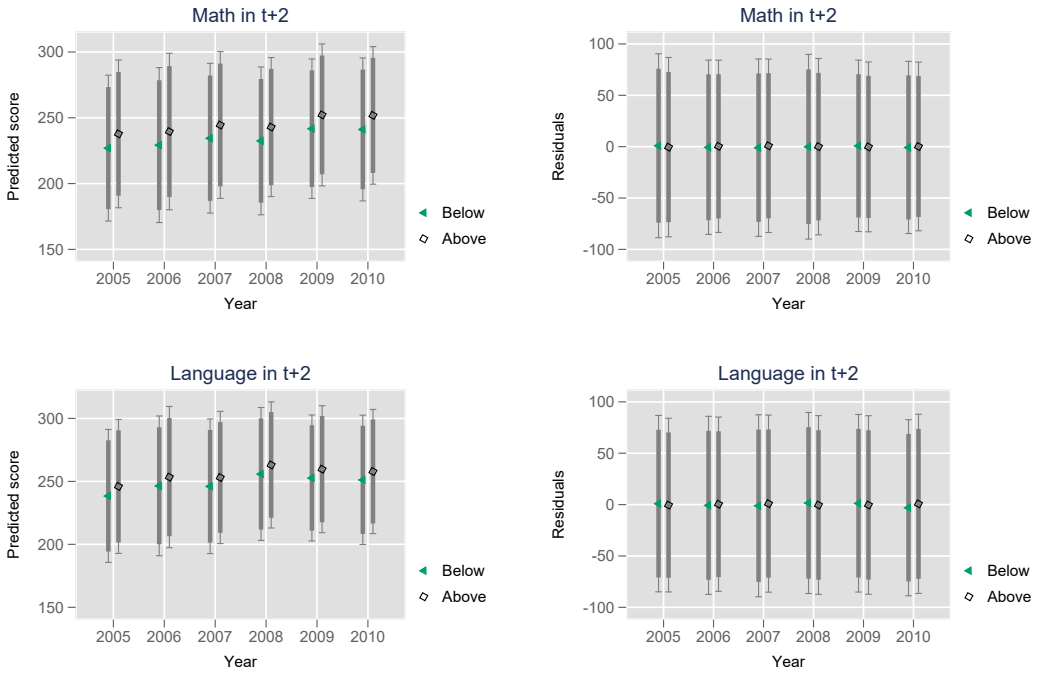
Taut, S., M. V. Santelices, C. Araya, and J. Manzi (2011, December). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. *Studies in Educational Evaluation 37*(4), 218–229.

Taylor, E. S. and J. H. Tyler (2012). The effect of evaluation on teacher performance. *The American Economic Review 102*(7), 3628–3651.

Appendix

Figures

Figure A1: (Absence of) Differential Trends During the Pre-Policy Period



Notes: This figure investigates differential trends in test-scores, during the pre-policy period. Left panels show predicted scores; right panels show residuals. Predicted scores and residuals stem from a reduced form regression, as shown in Equation 1. Top panels refer to math outcomes in year $t+2$; bottom panels refer to language outcomes in year $t+2$. "Below" refers to teachers qualifying for a "basic" rating (in year t); "above" refers to teachers qualifying for a better rating (in year t).

Tables

Table A1: Sample Characteristics and Validity Checks

	2005-2010		2011-13		
	Below	Above	Below	Above	DD
Teacher Baseline Characteristics					
t+1: Gender: Female	0.72	0.8	0.78	0.86	0.01 (0.03)
t+1: Contract hours	39.03	38.49	37.92	37.46	-0.12 (0.50)
t+1: Works in yet another school	0.08	0.07	0.04	0.02	0.01 (0.02)
t+1: Years in service	20.29	19.29	14.39	14.55	-0.53 (0.76)
t+1: n	1,972	3,700	296	1,665	7,633
t+1: School's baseline reading score†	241.02	247.14	248.88	253.48	-0.84 (1.50)
t+1: School's baseline math score†	229.19	235.7	237.88	244.72	-0.61 (1.72)
t+1: n	1,439	2,953	243	1,526	6,161
t+3: Gender: Female	0.75	0.83	0.85	0.88	0.04 (0.03)
t+3: Contract hours	38.67	38.24	36.8	36.69	-0.07 (0.63)
t+3: Works in yet another school	0.06	0.07	0.03	0.03	0.01 (0.02)
t+3: Years in service	20.12	18.42	12.16	13.89	-2.44 (0.95)**
t+3: n	1,795	3,402	185	795	6,177
t+3: School's baseline reading score†	241.15	247.64	251.75	255.94	1.40 (1.79)
t+3: School's baseline math score†	229.89	235.95	242.43	247.84	1.29 (2.06)
t+3: n	1,330	2,786	171	764	5,051
Student Baseline Characteristics					
t+1: GPA	5.77	5.83	5.73	5.79	0.01 (0.02)
t+1: Repeated in baseline year	0.03	0.02	0.02	0.02	-0.00 (0.00)
t+1: Attendance	92.78	93.12	90.99	91.9	-0.34 (0.29)
t+1: Gender: Female	0.48	0.49	0.48	0.49	-0.01 (0.01)
t+1: n	39,256	81,126	5,429	36,330	162,141
t+1: Household income (pesos)††	264,640	267,643	301,820	307,547	-7,044.85 (9,595.34)
t+1: Mother's edu. (years)††	9.84	10.02	10.4	10.88	-0.11 (0.12)
t+1: n	33,873	70,040	4,670	32,061	140,644
t+3: GPA	6.01	6.06	5.96	6	0.00 (0.03)
t+3: Repeated in baseline year	0.04	0.03	0.04	0.04	-0.00 (0.00)
t+3: Attendance	92.24	92.46	90.26	91.06	-0.11 (0.30)
t+3: Gender: Female	0.48	0.48	0.48	0.49	-0.01 (0.01)
t+3: n	35,051	74,699	3,998	18,817	132,565
t+3: Household income (pesos)††	279,234.67	288,412	358,775.1	339,968.9	16,032.12 (18,434.46)
t+3: Mother's edu. (years)††	10	10.22	11.36	11.36	0.01 (0.15)
t+3: n	30,140	64,254	3,459	16,600	114,453

Notes: "Teachers" include all unique year-teacher observations and may thus repeatedly include individual teachers over time. "Students" include all unique year-teacher-student observations and may thus include up to two observations per student and year (if math and reading are taught by different teachers, in a given year). "Below" and "Above" refer to teachers below or above the cut-off, respectively. t refers to the year of the initial evaluation. All variables measured in t, if not denoted otherwise. † denotes variables available for fewer observations (and not included as covariates). †† denotes variables measured at follow-up (and not included as covariates). Note that the 2013 sample is not followed up in $t+3$. "DD" refers to a difference-in-difference estimate as described in Section 4 (excluding control variables but including commune-level fixed effects). Standard errors in parentheses. For student-level characteristics, standard errors are clustered at the year-teacher level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.