Evaluating Teacher Evaluation:

Evidence From Chile

Andreas de Barros*

29 October 2018

Draft – Do Not Circulate

Abstract

This study investigates the causal effects of repeat performance evaluations, under Chile's national teacher evaluation system. The paper's identification strategy exploits a recent change in the evaluation assignment mechanism, together with a discontinuity in the system's scoring mechanism. Jointly, these two factors provide plausibly exogenous variation in which teachers get newly evaluated, after two years. The study's main results suggest that student learning, teacher beliefs and teaching behaviors remain unaffected by a teacher's re-evaluation, both in the year of the evaluation and in the year thereafter. The article confirms that these findings are not driven by a teacher's level of work experience, by student sorting, or by systematic attrition.

*Harvard Graduate School of Education. adebarros@g.harvard.edu. For support and comments, I would like to thank Felipe Barrera-Osorio, Olivia Chi, José Ignacio Cuesta, Melissa Dell, David Deming, Pierre de Galbert, Kathryn Gonzalez, Heather Hill, Francisco Lagos, Anne Lamb, Cristián Larroulet, María Lombardi, Karthik Muralidharan, Abhijeet Singh, Ugo Troiano, and Martin West. I thank the Chilean Agencia de Calidad de la Educación and the Ministry of Education for making data available. The usual disclaimer applies.

1 Introduction

While claims that performance evaluations lead to improved teacher productivity are controversial (Darling-Hammond et al., 2012), there is surprisingly little empirical research to inform related debates.[1] Moreover, there are strong disagreements on whether evaluations should be solely based on teachers' ability to raise test scores, with frequent demands for more comprehensive, "formative" evaluation systems (see Grissom and Youngs, 2016). Proponents of such formative approaches frequently refer to Chile's evaluation system as a best-practice example: For instance, a recent World Bank report concludes that teacher evaluation is "the essential backbone for a high-performing education system" and 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 paper answers three main questions. First and foremost, do repeat evaluations lead to improvements in teacher productivity, as measured by student learning? Second, how are potential mechanisms affected that are expected to enhance student performance? A key characteristic of formative evaluations is their objective to improve instructional practices and to affect commonly held beliefs among teachers. I therefore investigate intermediary effects of evaluations, on teaching behaviors and on teacher beliefs. 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

[1] Taylor and Tyler (2012) provide the so far only rigorous evaluation. I review their study and related literature in the subsequent section.

[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".

first five to ten years on the job (for a recent overview, see Kraft et al. (2018)). Hence, I assess heterogeneous effects of evaluations, by teachers' level of work experience.

The paper's 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 publicly funded 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 4^{th}-graders in mathematics and reading. Chile's teacher evaluation system uses recommended cut-off scores in order to categorize a teacher's continuous performance score into its four, discrete performance categories. 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 exploits this variation across time, together with the discontinuity in the evaluation system's underlying scoring mechanism. I show that, although the newly introduced law is imperfectly observed, it sharply increased a "basic" teacher's likelihood to be newly evaluated after two years.

The study's analytic findings rest on a difference-in-difference estimation strategy, which recovers the effect of teacher re-evaluations for those teachers who are assigned to be re-evaluated and comply with their assignment (the "treatment-on-the-treated" or "ToT" effect). To confirm their robustness, these estimates are moreover compared to those using a novel difference-in-discontinuities ("dif-in-disc") estimator. For the purpose of this comparison, I extend a previously proposed difference-in-discontinuities ("dif-in-disc") estimator (Grembi et al., 2012) to the given case, in which the assignment mechanism is "fuzzy" (rather than "sharp"). Intuitively, this latter strategy thus takes the difference between a fuzzy regression-discontinuity (RD) estimate after 2011 and a respective estimate for the period before 2011. The choice of both analytic strategies is motivated by (and overcomes)

two limitations of a simple RD estimator. First, I show that these approaches account for observable manipulation close to the cut-off score. Second, both econometric strategies take into account that the same cut-off score is used for additional national programs (both before and after the policy change), such as a national teacher incentive scheme and professional development programs. In additional analyses, the article also confirms that its results are not driven by student sorting or by systematic attrition.

Independent of the choice of estimation strategy, the study's main results suggest that student learning, teacher beliefs and teaching behaviors remain unaffected by a teacher's re-evaluation, both in the year of the evaluation and in the year thereafter. These results do not differ for teachers with fewer years of work experience. These findings provide first evidence on the impact of repeat-evaluations (in the given case, evaluations of teachers who have been evaluated just two years prior). To my best knowledge, this study furthermore provides the first quasi-experimental assessment of a national workforce evaluation system's effects on worker productivity. Thus, it may also offer interesting insights for public human resource management beyond the education sector.

The remainder of this paper is organized as follows. The next section (Section 2) describes theoretical considerations and debates the study's relationship to previous research. Section 3 gives a short introduction to Chile's teacher evaluation system. This is followed by a description of the paper's data sources, in Section 4.1. A subsequent section (Section 4.2) provides summary statistics for the analytic sample, and scrutinizes the study's internal validity. Section 4.3 introduces the proposed estimation framework, including information on the study's fuzzy difference-in-difference and dif-in-disc approaches. Section 5 presents results and Section 6 concludes.

2 Theoretical Considerations and Previous Research

From a theoretical viewpoint, there are no clear expectations considering whether, if at all, teacher evaluations have a positive or negative impact on teacher productivity and student achievement. A short review of five theoretical lenses illustrates this point. To begin, *principal agent theory* may 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). Similarly, 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. 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. Further, a *professionalization argument* hypothesizes that student learning may be improved if increases in teacher evaluation and accountability allow teaching to graduate from a "second grade" to a "full" profession (Johnson and Fiarman, 2012; Mehta, 2013). Yet, Mehta (*ibid.*) argues that professionalization may also be hindered if such accountability is to superiors only, rather than to colleagues. Finally, critics of teacher evaluations point to rather practical concerns and to an *opportunity-cost argument*, suggesting that teacher evaluations may take up scarce financial resources and teachers' work-time (*cf.* Taut et al., 2011).

At the same time, there is a dearth of empirical research on the effects of teacher evaluations on student learning. More broadly, other studies have investigated the joint impact of teacher evaluation and performance incentives (Daley and Kim, 2010; Dee and Wyckoff, 2015; Glazerman and Seifullah, 2012). However, these studies are not able to disentangle the effect of teacher evaluation from the respective incentive schemes. This observation also

holds true for an even wider body of research concerned with interventions related to teacher accountability and performance based management (Hanushek and Raymond, 2005; Dee and Jacob, 2009; Gerrish, 2014). Further, the present paper relates to studies focusing on programs that feature teacher evaluation or feedback on teaching practices as a key component, including peer collaboration such as lesson study and instructional rounds (Gersten et al., 2010; Louis and Marks, 1998), as well as tutoring, mentoring, and coaching (Bowman and McCormick, 2000; Murray et al., 2009; Johnson and Fiarman, 2012; Allen et al., 2011). Yet, related findings are mixed and studies are commonly criticized for "not meet[ing] the standards of rigorous research" (Cornett and Knight, 2009, 192).

As an exception to this lack of rigorous evidence, Taylor and Tyler (2012, henceforth: T&T) provide a quasi-experimental analysis in Cincinnati Public Schools, using a sample of 105 mid-career teachers. Taylor and Tyler (*ibid.*) apply a teacher fixed-effects strategy to exploit variation in the timing of each teacher's evaluation, over a time period of six years. In summary, the authors find that math test scores of students whose teacher was evaluated in the previous year increased by about ten percent of a standard deviation. The authors also conclude that this effect is greater for teachers whose previous performance (as judged by their students' test scores in the year before) was lower. T&T do not find an effect on reading scores. As discussed in the next section, Chile's system shares several characteristics with Cincinnati's system, such as its focus on a formative evaluation through peer-reviews and classroom observations. However, the present paper adds to Taylor and Tyler's work in at least three distinct ways. First, the present study analyzes the effects of a fully established evaluation system at national scale, rather than a newly introduced system covering a comparatively small sample of teachers, in a single district. Secondly, Taylor and Tyler (2012, 3647) point out that their findings may be due to the fact that teachers received their very first evaluation after 8 to 17 years on the job, concluding that "the effect resulting from each subsequent year's evaluation might well be smaller". The present study examines exactly

such a case as all teachers in my sample had been evaluated two years prior. Third, T&T's analyses of mechanisms is limited. In contrast, this study investigates an evaluation's effects on intermediate factors such as a teacher's beliefs, attitudes, and teaching behaviors.

3 Teacher Evaluation in Chile

This section provides a short overview of Chile's teacher performance evaluation system, Docentemás, focusing on those characteristics that provide the study's source of exogenous variation.[4] Docentemás was introduced in 2003 as a standards-based assessment system that is tied to Chile's common quality Framework of Good Teaching (Marco para la Buena Enseñanza, MBE).[5] In 2005, participation became mandatory, for all public schools in the country.[6] The 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%).[7] The latter, in turn 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 (outstanding, competent, basic, and unsatisfactory). 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

[4]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.

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

[6]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".

[7]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%).

by a Municipal Evaluation Commission before it becomes final (modifications occur in approximately five percent of cases). Once the rating category has been decided upon, detailed, written feedback is provided to teachers, schools, and municipal education authorities. 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, and further reports are distributed to other parties in April.[8] Interestingly, results for the largest evaluation component (anonymous ratings of the teacher performance portfolio) only become available *after* the remaining three components have been scored. 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.

In 2011, a new law (Ley 20.501) introduced changes to the consequences of a teacher's performance rating. Generally, municipal teachers are required to be evaluated at least once every four years.[9] Yet, under the new law, teachers rated as "basic" have to be re-evaluated after two years.[10] The law came into effect in 2011, but it did not affect teachers retroactively, based on their previous performance ratings. For "basic" teachers, the new law also did not result in other changes – whether with respect to their job security, their access to incentive schemes, or their professional development, for example.

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

[9]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.

[10]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.

In terms of teacher turn-over, since 2011, a teacher with a "basic" rating has to leave the system if their rating does not improve in the next two assessments. Note however, that 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.

The same distinction in ratings – "basic" vs "competent" and "outstanding" – is not only used to trigger a teacher's renewed evaluation. Teachers in the top two categories also receive access to a rewards and incentive scheme (the Program for the Variable Individual Performance Allowance; in short: AVDI).[11] 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 (Professional Development Plans; in short: PSPs) before their next evaluation takes place.[12] Yet, crucially, assignment mechanisms for these programs did not change with the 2011 law.

Therefore, in the three-year period after their initial evaluation, 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 performance score suggested a "basic" rating (inducing

[11] 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).

[12] These 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.

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 Research Design

4.1 Data

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, and results on standardized test scores for all Chilean 4^{th}-graders (in mathematics and reading). More precisely, I combine data from five different sources.[13] The first data source is the "Ideoneidad Docente" data-base (available for the years 2003 to 2015), which is maintained by Chile's Ministry of Education. The data-set 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).

The study's second 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.[14] Third, 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

[13]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.

[14]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.

performance score, and her rating on each of the four sub-categories. 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). For the time-span of the given analysis (of teachers' "baseline" evaluations, *i.e.* from 2005 to 2013), this renders a data-set of 30,293,522 student-by-year observations. Finally, student learning outcomes are measured using Simce exam scores.[15] 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 language in every consecutive year since 2005.[16] Simce data-sets include information on the student's gender, school, and class, and additional student demographics (her mother's highest level of education and her 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.

This paper investigates mechanisms by including information from detailed, yearly questionnaires covering teacher beliefs and behaviors, as reported by parents, teachers, and students. For each year, Simce data-sets are complemented by a student, a teacher, and a parent survey. In particular, for each year from 2005 to 2014, questionnaires ask teachers about their confidence in students' future educational attainment.[17] Moreover, for the years 2012 through 2015, I construct an index of student-reported teaching effort. More specifically, I calculate a

[15]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.

[16]See Figure A1, in the Appendix, for more detailed information on which subjects were tested in a given grade and year (in Spanish).

[17]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.

simple average over six items that seek to capture a teacher's classroom behaviors.[18] Finally, for 2011 to 2014, I use a measure that asks parents to rate the extent to which the head teacher cares about her students.[19] As there is no head teacher identifier, in my analysis of mechanisms, I assume that in fourth grade, a teacher is considered the head teacher if she teaches both math and language.[20] Figure 1 explores whether each of these three measures are correlated with student learning. Each panel shows the linear fit from a simple bivariate regression of math and language Simce scores on the three indicators of teacher beliefs and behaviors (as well as the respective 95 percent confidence interval).[21] Each panel also includes binned scatter plots. In summary, Figure 1 suggests that the three mechanisms are strongly correlated with student learning (in both mathematics and language), and that this relationship is approximately linear.

[Figure 1 about here]

4.2 Sample Characteristics and Threats to Internal Validity

Table 1 provides summary statistics for the study's sample of teachers at "baseline" (the year of their initial evaluation). 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.[22] I also drop the small share of approximately 1.8 percent of

[18]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).

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

[20]Approximately, 90 percent of the study's sample of 4th-graders have the same math and Spanish-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.

[21]I use data for a sample of teachers and their students two years after a teacher's initial evaluation (i.e., the year in which the re-evaluation is expected to occur). Graphs for one year and three years post-evaluation are available upon request and provide virtually identical results.

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

teachers with a performance rating that would have suggested an "unsatisfactory" rating.[23] This renders a sample of 29,093 teachers who were initially evaluated in 2005-2013, of which 8,363 teachers were initially evaluated after the law was introduced, in 2011, 2012, or 2013.[24]

I investigate (and present evidence against) three potential threats to the study's internal validity: differential attrition based on a teacher's assignment status, imbalance of observable teacher characteristics across "treated" teachers and their comparison group, and potential sorting of students to (or away from) teachers who are induced to be re-evaluated. As shown in Table 1's top-most panel, there is no significant difference in the percentage of teachers who are subsequently observed in the paper's analytic samples of 4th-grade Simce teachers (one, two, or three years past the assignment), suggesting that there is no systematic attrition due to teachers' performance on the evaluation (whether before or after the policy change, or with respect to the difference-in-differences). Of the initially evaluated teachers, 7,037, 6,883, and 5,721 teachers are observed in the three years after, teaching a total of 159,134, 153,724, and 129,827 4th-grade students, respectively.[25]

Table 1's penultimate column follows the paper's preferred difference-in-difference strategy (as described in Section 4.3 below, with the exclusion of covariates) and reports whether the introduction of the law coincided with imbalance in the reported baseline variables. In robustness checks, I also reduce the sample for a comparison of teachers within a narrow band of 0.2 performance score points above and below the "basic" vs "competent" cut-off.[26] As teacher evaluations are at least partly implemented at the commune-level, these and

[23]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.

[24]More precisely, these numbers reflect year-teacher observations.

[25]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 $t + 3$.

[26]The following section (Section 4.3) discusses bandwidth selection and introduces the respective difference-in-discontinuities estimator.

all remaining analyses include commune-fixed effects. For the three samples of 4th-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 (in a given year). At baseline, I also find no differences in teachers' average school-level Simce score (whether in 4th-grade reading or math). As an exception, I find that, with the introduction of the new policy, teachers in the sample for $t + 2$ who were induced to be re-evaluated taught slightly more contract hours at baseline (an additional 0.85 hours, according to the difference-in-difference estimate). This difference is marginally significant, and is only found in the second of the three follow-up samples. Furthermore, in the post-period, assigned teachers in the sample for $t + 3$ were found to be slightly younger (2.5 years, significant at the 0.01 level) and with fewer years of teaching experience (2.2 years, significant at the 0.05 level). Again, these differences are not confirmed in the remaining two samples.

To investigate the potential of systematic sorting, Table 1 also includes descriptive statistics for teachers' 4th-grade Simce-taking students, in the three years post-assignment (yet measured at baseline). The table also presents information on student demographics (household income and the mother's highest level of education), but 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. Importantly, I find no support for the hypothesis that schools engage in sorting of students to (or away from) teachers that are assigned to be re-evaluated. None of the 15 tests point to differences in student characteristics (at the 0.1 level, and independent of the analytic strategy). 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.

[Table 1 about here]

4.3 Identification Strategy

The availability of a continuous assignment variable with a cut-off rule may – misleadingly – point to a simple regression-discontinuity (RD) strategy. This section begins by discussing the inappropriateness of such an approach and presents its preferred estimation framework thereafter. First, recall that, even with a balanced sample, 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 good as random at the threshold score, which implies that teachers are not able to manipulate their scores around this cut-off. Yet, as teachers, their peers, and principals determine 40 percent of the performance score (through self- and peer-evaluations, as well as reference-reports), this assumption seems questionable if teachers are "pushed over" the cut-off score.

In contrast, however, the paper's analysis finds the opposite: in both the pre- and post-policy periods, the density of teachers just below the cut-off score was slightly higher than expected (from a smooth density distribution), in comparison to teachers just above the cut-off score. The top-panel of Figure 2 shows McCrary (2008) 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 rejects the null of equal densities on both sides (at the 0.01 level, for both periods). 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[27], and smoothing over the histogram thereafter (separately,

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

for both sides of the breakpoint).[28] 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 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.

[Figure 2 about here]

This study therefore proposes a fuzzy difference-in-difference ("fuzzy DD") estimation strategy. In summary, 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. My proposed estimator moreover accounts for the fact that the law is not observed perfectly, i.e. that the jump in the probability of getting evaluated two years after is "fuzzy". More formally, I estimate a two-stage least squares (2SLS) regression, whose reduced form is given in equation 1, as follows.[29]

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

Reduced-form (RF) equation 1 refers to teacher j, initially evaluated in year t. Here, Y denotes 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 the breakpoint in the post-policy period (reflecting assignment to re-evaluation), for teacher scores θ. Γ_t captures year fixed effects; Ω captures commune fixed effects (I omit a commune

[28] 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.

[29] 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.

subscript throughout); $\boldsymbol{X}$ is a vector of teacher and student characteristics in baseline year t (as shown in Table 1).[30] The respective first-stage (FS) equation (not shown) is equivalent to equation 1, but now the dependent variable $Y_{j(t+2)}$ reflects a teacher's re-evaluation in $t+2$, 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 excludes student characteristics.

In robustness checks, the paper compares its difference-in-differences results to those of a fuzzy difference-in-discontinuities ("fuzzy dif-in-disc") estimation strategy.[31] This approach leverages that, in addition to the counter-factual continuity of differences, *in any year*, treatment assignment can moreover be expected to be as good as random, within a narrow bandwidth around the assignment breakpoint.[32] As common for RD estimates, the paper's dif-in-disc strategy moreover controls for a trend in the forcing variable[33], and allows for this trend to differ on either side of the cutoff as well as after the introduction of the new law. The respective reduced form is given in equation 2, as follows:

$$\begin{aligned} Y_{j(t+x)i} =& \beta_{RF0} + \beta_{RF1} T_{jt} + \beta_{RF2} TxPost_{jt} + \beta_{RF3} Tx\theta_{jt} + \beta_{RF4} Postx\theta_{jt} \\ &+ \beta_{RF5} \theta_{jt} + \Gamma_t + \Omega + \boldsymbol{X}_{jti} + \epsilon_{j(t+x)i} \end{aligned} \tag{2}$$

$Postx\theta$ indicates the interaction term between a post-period indicator and the forcing variable; $Tx\theta$ denotes its interaction with an indicator of being below the threshold. All remaining notation is as in equation 1, above. With both estimation approaches, the coefficient of main interest is β_2. In the remainder of the paper, all reported effect sizes (and their stan-

[30] 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).

[31] My approach extends prior work by Veronica Grembi, Tommaso Nannicini and Ugo Troiano (2012), who propose a difference-in-discontinuities estimator for a case in which the discontinuity is "sharp".

[32] As shown in the above analysis of smooth covariates, even if the respective RD assumptions are violated, the difference-in-difference of RD estimates, or "difference-in-discontinuity" estimates, nevertheless hold.

[33] I follow Gelman and Imbens (2014), confirm the validity of a linear trend by visual inspection of binned scatter plots (not shown), and do not include higher-order polynomials.

dard errors) represent the Wald estimate, where $\beta_{Wald2} = \beta_{RF2}/\beta_{FS2}$. Given the (complete lack of) re-evaluation rates in the pre-period and near-zero rates for the post-period comparison groups (see below), β_{Wald2} is interpreted as a Treatment-on-the-Treated (ToT) effect.

Finally, in the paper's analysis of heterogeneous effects, I furthermore include interactions between the 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.[34]

5 Analysis and Results

5.1 First Stage Results

Figure 3 provides graphical 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, e.g. 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 48 percent; in contrast, the percentage remains close to zero once the threshold is crossed (ranging from 0 to 3 percent). Note that there is great variance with respect to compliance (or the level of "fuzziness") among the assigned

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

teachers. Yet, in addition to the visual evidence above, 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 above 20, for all three samples ($t+1$, $t+2$, and $t+3$).

[Figure 3 about here]

5.2 Main Results

I calculate the study's two-stage least-squares Wald estimates within a common bootstrap procedure (with 750 replications) and obtain clustered standard errors by blocking resamples at the classroom level. I repeat this procedure for both subjects and for the three points in time after a teacher's assignment (*i.e.* $x = 1$, $x = 2$, and $x = 3$). Table 2 shows the study's main difference-in-difference results. Figure A2 in the Appendix summarizes the same results graphically. 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.[35] 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, 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, *i.e.* one year after the "treatment" was assigned

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

$(t+1)$. 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, with the exception of language teachers in year $t+3$, who react less negatively to re-evaluations when compared to their colleagues with fewer years of experience (in expectation, approximately 0.1 standard deviations, for a teacher with 10 years more experience, significant at the 0.1 level). Note that this pattern is not confirmed with respect to results in mathematics.

Figure 4 investigates whether these results are robust to the estimation strategy (*i.e.* to varying bandwidths and to the inclusion of a trend in the forcing variable). I repeat the above estimations with a very narrow choice (0.1), and thereafter increase the bandwidth in increments of 0.1 points, up to 0.4. Starting from the left, each dot refers to the corresponding estimate for the main coefficient of interest (β_{Wald2}) for increasing bandwidths; the right-most dot reflects the difference-in-difference estimate (as shown in Table 2's columns 1, 3, and 5. 95 percent confidence intervals are given in capped gray lines; 90 percent confidence intervals are shown in thick gray lines. As can be seen for the left-most estimates, independent of the subject (math or language), error bands become extremely large for a bandwidth of 0.1. However, as expected, the error bands narrow with increased bandwidth and the point estimates approach those of the difference-in-difference estimator. None of the reported 90 percent confidence intervals exclude the null. 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.

[Figure 4 about here]

5.3 Mechanisms

Table 3 reports on the effects on head-teachers' (student-reported) teaching behaviors, head-teachers' (parent-reported) levels of caring, and head-teachers' (self-reported) beliefs in their students' future educational attainment. All measures are standardized. As with the analysis of main results, all models include a vector of baseline teacher and student characteristics, commune and year fixed effects, and standard errors are clustered at the classroom level. Figure A3 in the Appendix presents the same results graphically.

[Table 3 about here]

In summary, whether detrimental or desirable, I find no evidence for effects of a teacher's re-evaluation on teaching behaviors, teachers' level of caring, or teachers' beliefs in students' future educational attainment (at the 0.1 level), whether in the year of or in the year after a teacher's re-evaluation. Note that, given the low number of observed years, coefficients on teachers beliefs are imprecisely estimated. In contrast, for the year prior to the re-evaluation ($t + 1$), I find that evaluated headteachers reduced their level of caring by .27 standard deviations (significant at the 0.05 level). None of the results differ by teachers' level of work experience.

6 Conclusion

This study intends to offer a contribution to discussions on the effects of performance evaluations on teacher productivity and children's educational achievement. In doing so, to my best knowledge, this study provides only the second rigorous assessment of this relationship and the first analysis of teacher evaluations under a well-established evaluation system that operates at national scale. In summary, I cannot conclude that Chile's repeat performance evaluations lead to substantial gains in student achievement, one year after a teacher's re-evaluation, in both math and reading. However, for the year of the re-evaluation, I also do

not find any negative effects on student learning. At the same time, I find that concerns about additional detrimental effects may be at least partly warranted: The paper's results suggest that re-evaluations led to decreases in teachers' level of caring, but only in the year prior to the re-evaluation. I do not find support for additional negative effects, as measured by teachers' beliefs in their students future educational attainment or their teaching practices.

Of course, this study is not without its challenges. For example, only about a quarter of elementary teachers are observed teaching a fourth-grade class. In addition, given the comparatively smaller number of cohorts with information on teachers' beliefs, related standard errors are quite large. At the same time, the study has already investigated several of these or related shortcomings. For instance, given that most Chilean elementary schools teach four grades, it is not surprising that only a quarter of teachers are observed in a given year. Importantly, I find no evidence for a relationship between a teacher's assignment to re-evaluations and whether the teacher teaches a fourth-grade math or language class. In addition, by relying on detailed, administrative records for the universe of Chilean students, as well as on additional demographic information, the study is able to rule out concerns about systematic sorting of students away from (or towards) teachers. Lastly, the article shows that its findings are robust to the choice of alternative econometric strategies – that of a "more conventional" difference-in-difference estimator, or that of an even more conservative, yet lower-powered, difference-in-discontinuity estimator (with a number of varying bandwidths).

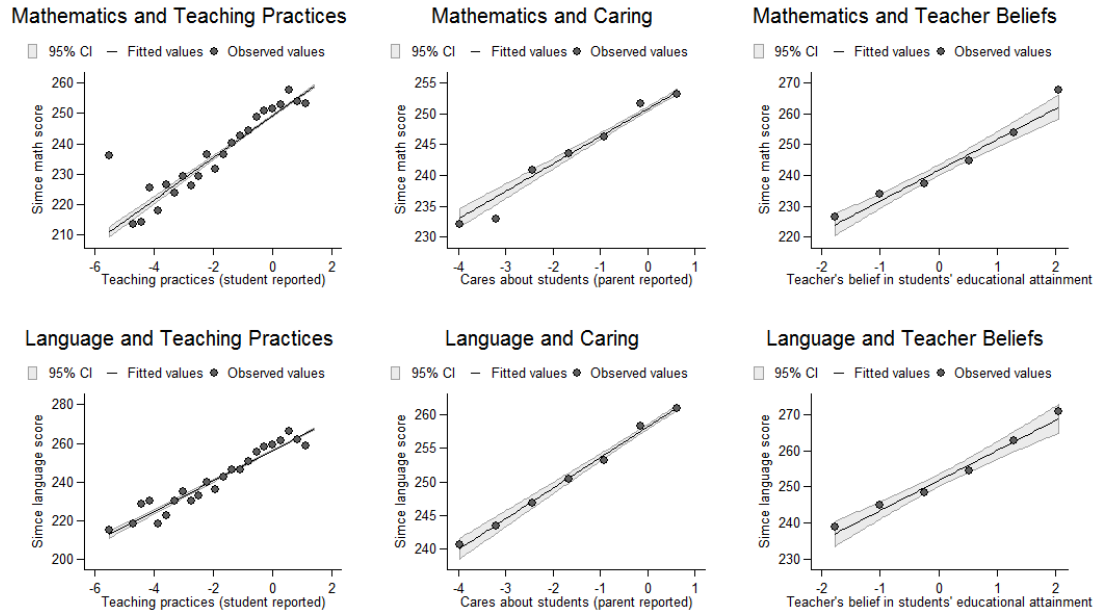
In evaluating the impact of "Docentemás", this study aims to provide first evidence on the effects of a comprehensive, standards-based teacher evaluation system that has recently been described as a role model for other countries (Bruns and Luque, 2014). Further, as discussed by Taut et al. (2011), Chilean 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. I am confident that such debates can be fostered by providing sound evidence on the effects of Chile's evaluation system on teacher productivity (as measured through student performance), teacher beliefs, and teaching behaviors.

Figures and Tables

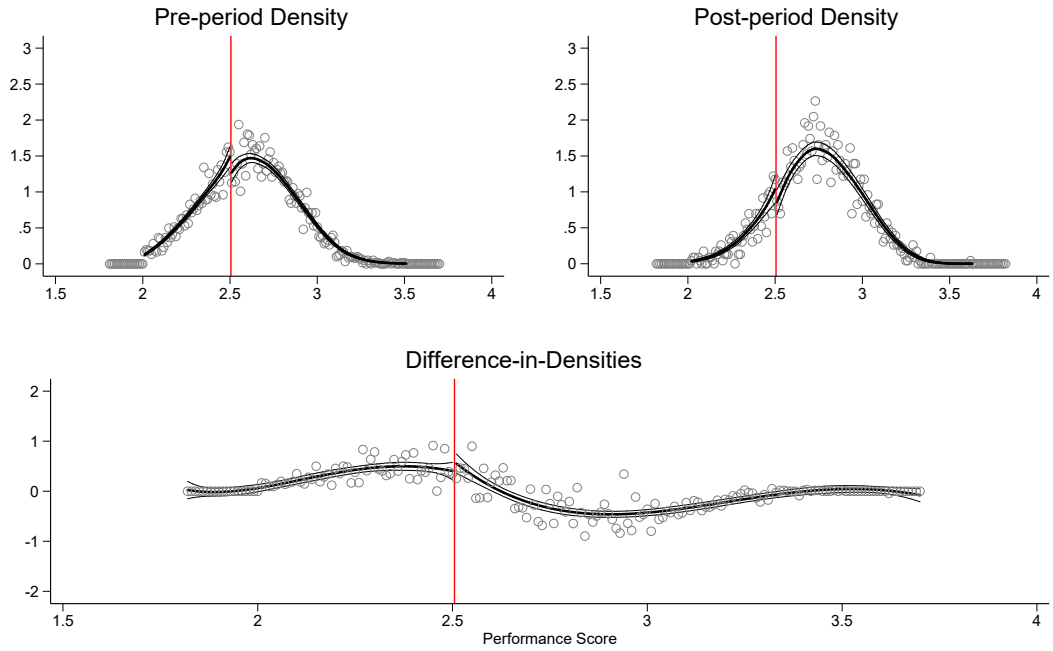
Figures

Figure 1: Bivariate Relationships Between Student Learning, Teacher Behaviors, and Teacher Beliefs



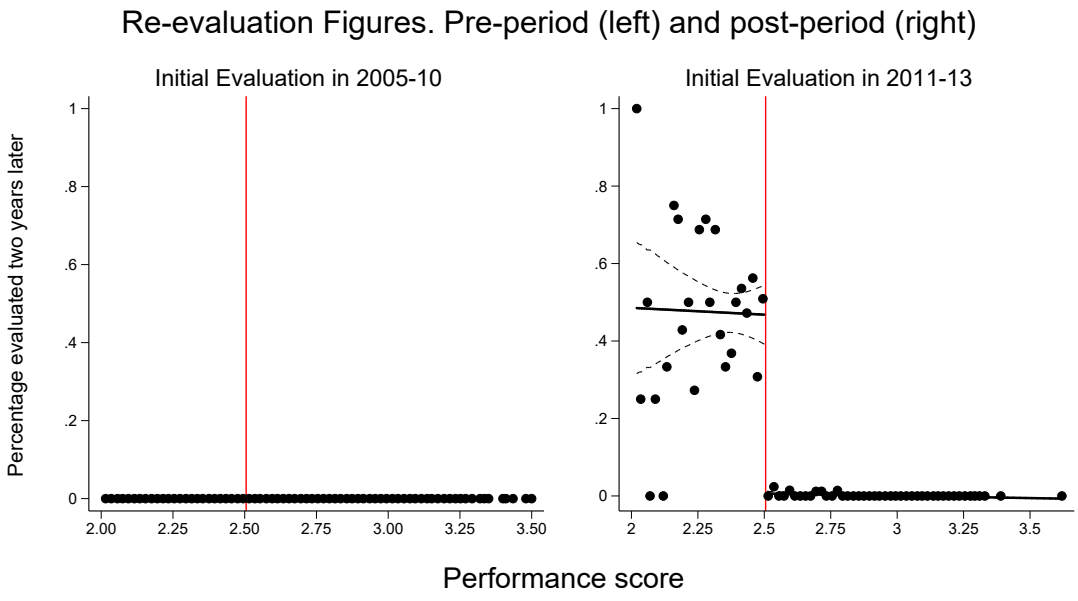
Note: Observed values show the mean Simce score within equal-sized bins of the predictor variable.
All increments used to construct bins for caring and beliefs. 25 bins used for teaching practices.
Fitted values and confidence intervals use all observations, i.e. not the binned averages. Sample: Headteachers in t+2.

Figure 2: McCrary Plots



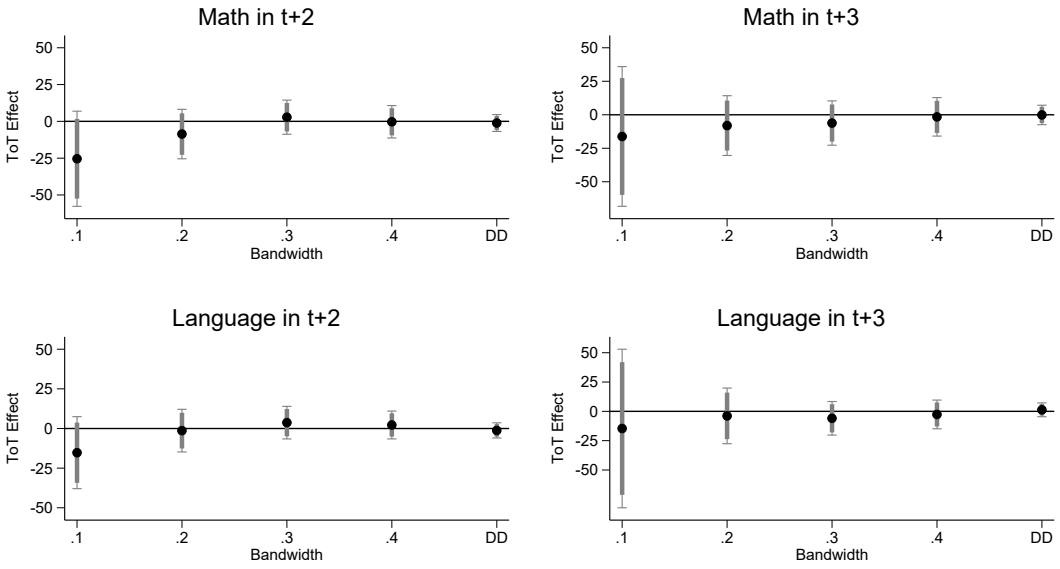
Note: Vertical lines indicate the recommended cutoff score.

Figure 3: 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 4: Sensitivity to Bandwidth Selection



Notes: Each dot refers to the ToT effect, estimated at varying bandwidths.
DD reflects the difference-in-difference estimation,
results for the remaining bandwidths reflect the fuzzy difference-in-discontinuities estimation.
Bootstrapped standard errors (750 replications), blocked at the classroom level.
95 percent confidence interval in capped gray lines. 90 percent confidence interval in thick gray lines.

Tables

Table 1: Sample Characteristics and Validity Checks

	2005-2010			2011-2013			DD	Dif-in-Disc
	Below	Above	RD	Below	Above	RD		
Attrition								
In sample in t+1	0.23	0.26	-0.00 (0.02)	0.19	0.23	-0.04 (0.03)	-0.02 (0.01)	-0.03 (0.03)
In sample in t+2	0.22	0.25	0.00 (0.02)	0.18	0.23	-0.01 (0.03)	-0.02 (0.01)	-0.03 (0.03)
In sample in t+3	0.21	0.24	0.02 (0.02)	0.12	0.11	-0.01 (0.02)	0.00 (0.01)	-0.03 (0.03)
n	7352	13378	10908	1420	6943	3503	29093	14411
Teacher Baseline Characteristics								
t+1: Gender: Female	0.75	0.83	-0.05 (0.03)	0.79	0.87	-0.04 (0.08)	0.00 (0.03)	0.07 (0.07)
t+1: Age	45.63	43.88	-0.50 (0.71)	41.34	40.27	-1.50 (2.01)	0.01 (0.67)	-1.14 (1.77)
t+1: Contract hours	38.41	38.06	0.01 (0.65)	37.57	37.28	2.05 (1.19)*	-0.01 (0.51)	0.57 (1.22)
t+1: Works in yet another school	0.09	0.07	-0.00 (0.02)	0.04	0.02	0.05 (0.04)	0.01 (0.02)	0.03 (0.04)
t+1: Years in service	20.08	19.16	-0.42 (0.89)	14.04	14.38	-1.67 (2.15)	-0.41 (0.79)	-1.86 (1.96)
t+1: School's baseline reading score	241.37	247.38	0.11 (1.67)	248.97	253.61	1.26 (4.80)	-1.07 (1.55)	4.08 (3.98)
t+1: School's baseline math score	228.22	235.38	-0.93 (1.93)	237.33	244.48	3.18 (5.21)	-0.06 (1.76)	6.26 (4.49)
t+1: n	1725	3433	2635	275	1604	749	7037	3384
t+2: Gender: Female	0.79	0.85	-0.01 (0.03)	0.79	0.88	-0.02 (0.07)	-0.04 (0.03)	0.01 (0.07)
t+2: Age	45.24	43.5	-0.87 (0.74)	40.67	40.75	2.55 (2.11)	-1.07 (0.69)	0.37 (1.79)
t+2: Contract hours	37.95	37.87	-0.41 (0.67)	37.81	37.08	1.66 (1.31)	0.85 (0.52)*	2.10 (1.25)*
t+2: Works in yet another school	0.07	0.06	0.03 (0.02)	0.04	0.03	0.00 (0.04)	0.01 (0.02)	-0.02 (0.04)
t+2: Years in service	19.54	18.54	-1.33 (0.94)	13.65	14.7	-0.31 (2.30)	-1.01 (0.82)	-0.22 (2.03)
t+2: School's baseline reading score	241.34	247.97	2.46 (1.84)	249.87	255.13	-2.06 (4.75)	-0.12 (1.59)	-1.60 (3.90)
t+2: School's baseline math score	229	235.64	2.52 (2.04)	238.53	246.46	2.38 (4.60)	-1.04 (1.78)	-0.86 (4.07)
t+2: n	1650	3379	2566	259	1595	729	6883	3295
t+3: Gender: Female	0.78	0.86	-0.01 (0.03)	0.85	0.89	-0.10 (0.10)	0.03 (0.03)	-0.07 (0.08)
t+3: Age	45.27	43.21	0.86 (0.79)	39.2	40.29	0.82 (2.73)	-2.46 (0.82)***	-3.05 (2.18)
t+3: Contract hours	38.15	37.81	-0.75 (0.67)	36.6	36.5	0.59 (1.26)	0.03 (0.63)	1.35 (1.23)
t+3: Works in yet another school	0.07	0.07	-0.02 (0.02)	0.03	0.03	0.05 (0.04)	0.01 (0.02)	0.06 (0.04)
t+3: Years in service	19.98	18.29	1.71 (0.96)*	12.31	13.81	0.36 (3.14)	-2.24 (0.98)**	-3.32 (2.57)
t+3: School's baseline reading score	240.97	247.81	-0.04 (1.84)	251.73	256.1	7.36 (6.29)	1.92 (1.89)	1.42 (4.94)
t+3: School's baseline math score	228.34	235.62	-1.04 (2.11)	242.14	247.96	11.97 (6.97)*	2.60 (2.16)	6.88 (5.45)
t+3: n	1586	3182	2464	178	775	454	5721	2918
Student Baseline Characteristics								
t+1: GPA	5.77	5.83	-0.00 (0.02)	5.73	5.79	-0.05 (0.04)	0.01 (0.02)	-0.01 (0.05)
t+1: Repeated in baseline year	0.02	0.02	-0.00 (0.00)	0.02	0.02	-0.01 (0.01)	-0.00 (0.00)	-0.00 (0.01)
t+1: Attendance	92.75	93.11	-0.47 (0.52)	91	91.92	-0.92 (0.62)	-0.33 (0.29)	-0.49 (0.70)
t+1: Gender: Female	1.48	1.49	-0.02 (0.02)	1.48	1.5	0.01 (0.03)	-0.01 (0.01)	0.02 (0.03)
t+1: n	38009	79941	117950	5298	35886	41184	159134	74740
t+1: Household income (pesos)†	267023.4	268866.3	-569.35 (8069.95)	301958.5	308048.5	-2.0e+04 (28443.50)	-7049.26 (8663.65)	-1.1e+04 (24889.17)
t+1: Mother's edu. (years)†	9.88	10.05	-0.12 (0.12)	10.42	10.9	-0.41 (0.27)	-0.11 (0.11)	-0.13 (0.27)
t+1: n	32842	69136	50995	4560	31763	13674	138301	64669
t+2: GPA	5.9	5.95	-0.02 (0.02)	5.87	5.91	-0.04 (0.04)	0.03 (0.02)	-0.00 (0.05)
t+2: Repeated in baseline year	0.03	0.03	-0.00 (0.00)	0.04	0.04	0.01 (0.01)	0.00 (0.00)	0.01 (0.01)
t+2: Attendance	92.63	92.97	-0.16 (0.25)	90.75	91.66	-0.64 (0.51)	-0.24 (0.24)	-0.35 (0.54)
t+2: Gender: Female	1.49	1.5	0.00 (0.01)	1.48	1.49	-0.02 (0.03)	-0.00 (0.01)	-0.02 (0.03)
t+2: n	35238	77830	113068	5271	35385	40656	153724	73606
t+2: Household income (pesos)†	267936.7	280121.4	13778.23 (11389.09)	344863.2	329243.3	-4.5e+04 (27215.05)	-2659.20 (10513.08)	-2.0e+04 (20329.13)
t+2: Mother's edu. (years)†	9.89	10.2	0.01 (0.13)	10.77	10.83	-0.31 (0.25)	-0.09 (0.11)	-0.15 (0.24)
t+2: n	28732	65133	47332	4570	31178	13868	129613	61200
t+3: GPA	6.02	6.06	0.03 (0.02)	5.97	6.01	0.05 (0.08)	0.00 (0.03)	-0.05 (0.07)
t+3: Repeated in baseline year	0.04	0.03	-0.01 (0.00)**	0.04	0.04	-0.01 (0.01)	-0.00 (0.00)	0.00 (0.01)
t+3: Attendance	92.25	92.47	-0.21 (0.26)	90.28	91.07	-0.21 (0.96)	-0.06 (0.31)	-0.77 (0.77)
t+3: Gender: Female	1.48	1.48	0.03 (0.01)**	1.48	1.49	-0.06 (0.04)	-0.02 (0.01)	-0.04 (0.04)
t+3: n	33970	73398	107368	3910	18549	22459	129827	66056
t+3: Household income (pesos)†	280724.3	289788.1	806.10 (11491.66)	359701.9	340602.8	23621.42 (45299.92)	16413.66 (16016.74)	-6350.94 (35440.64)
t+3: Mother's edu. (years)†	10.04	10.25	0.12 (0.12)	11.36	11.37	0.10 (0.38)	-0.01 (0.14)	-0.22 (0.32)
t+3: n	29286	63336	47815	3403	16449	9017	112474	56832

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. "RD" refers to a simple regression-discontinuity estimate. "DD" refers to the paper's preferred difference-in-difference estimate. "Dif-in-Disc" refers to the robustness check with a difference-in-discontinuity estimate. Both are as described further below, in Section 4.3, but exclude any control variables. RD and Dif-in-Disc estimates are calculated within a bandwidth of 0.2 around the cut-off and include linear trends of the forcing variable, which are allowed to vary on either side. t refers to the year of the initial evaluation. All variables measured in t, if not denoted otherwise. † denotes variables measured at follow-up (and not included as covariates). Note that the 2013 sample is not followed up in $t+3$. All values account for 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: Difference-in-difference Estimates: ToT Effects on Student Learning

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
β_{Wald2}	-1.887 (3.533)	-5 (7.655)	-1.109 (2.922)	-2.375 (4.295)	-.136 (3.707)	-.098 (6.08)
β_{Wald6}		.116 (.284)		.062 (.206)		-.071 (.263)
n (teachers)	6414	6335	6514	6430	4995	4924
n (students)	140161	140161	140058	140058	108255	108255
Language						
β_{Wald2}	-4.316 (4.181)	-7.41 (7.55)	-1.218 (2.442)	-3.801 (3.807)	1.31 (3.007)	-5.719 (4.85)
β_{Wald6}		.072 (.21)		.184 (.194)		.464 (.265)*
n (teachers)	6580	6495	6776	6686	5214	5129
n (students)	142522	142522	144754	144754	111156	111156

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 classroom level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Difference-in-difference Estimates: ToT Effects on Teaching Behaviors, Caring, Teacher Beliefs

	t+1		t+2		t+3	
	(1)	(2)	(3)	(4)	(5)	(6)
Practices						
β_{Wald2}	-.046 (.111)	.011 (.164)	.028 (.063)	.11 (.099)	.016 (.082)	-.041 (.119)
β_{Wald6}		-.003 (.007)		-.005 (.005)		.004 (.007)
n (teachers)	2148	2114	3153	3104	2322	2288
n (students)	44617	44617	65504	65504	49574	49574
Caring						
β_{Wald2}	-.272 (.112)**	-.298 (.16)*	.044 (.084)	.098 (.148)	.013 (.081)	.091 (.135)
β_{Wald6}		0 (.007)		-.002 (.007)		-.004 (.007)
n (teachers)	1916	1902	2127	2108	1780	1746
n (parents)	40305	40305	44127	44127	38551	38551
Beliefs						
β_{Wald2}	.024 (.197)	-.228 (.287)	.027 (.165)	.133 (.26)	-.175 (.272)	-.005 (.433)
β_{Wald6}		.015 (.012)		-.006 (.01)		-.01 (.027)
n (teachers)	5345	5282	4673	4609	3681	3631

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 classroom level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

References

Agencia de Calidad de la Educación (2015). Calendario de Evaluaciones.

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.

Bowman, C. L. and S. McCormick (2000). Comparison of peer coaching versus traditional supervision effects. *The Journal of Educational Research 93*(4), 256–261.

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.

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

Cornett, J. and J. Knight (2009). Research on coaching. In J. Knight (Ed.), *Coaching: Approaches and Perspectives*, pp. 192–216. Thousand Oaks: Corwin Press.

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.

Daley, G. and L. Kim (2010). A Teacher Evaluation System That Works. Technical report, National Institute for Excellence in Teaching (NIET), Santa Monica.

Darling-Hammond, L., A. Amrein-Beardsley, E. Haertel, and J. Rothstein (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 8–15.

Dee, T. and B. Jacob (2009, November). The Impact of No Child Left Behind on Student Achievement. Working Paper 15531, National Bureau of Economic Research, Cambridge.

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

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

Gelman, A. and G. Imbens (2014, August). Why High-order Polynomials Should not be Used in Regression Discontinuity Designs. Working Paper w20405, Social Science Research Network, Rochester, NY.

Gerrish, E. (2014, September). *The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis*. Dissertation, Indiana University, Indianapolis.

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.

Glazerman, S. and A. Seifullah (2012, March). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report. Technical report, Mathematica Policy Research, Inc., Princeton.

Grembi, V., T. Nannicini, and U. Troiano (2012, October). Policy Responses to Fiscal Restraints: A Difference-in-Discontinuities Design. Working Paper 6952, IZA.

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

Hanushek, E. A. and M. E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of policy analysis and management 24*(2), 297–327.

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

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.

Kraft, M. A., J. P. Papay, and O. L. Chi (2018, August). Teacher skill development: Evidence from performance ratings by principals. Working Paper, Brown University, Providence, RI.

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.

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.

Murray, S., X. Ma, and J. Mazur (2009). Effects of peer coaching on teachers' collaborative interactions and students' mathematics achievement. *The Journal of Educational Research 102*(3), 203–212.

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

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.

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

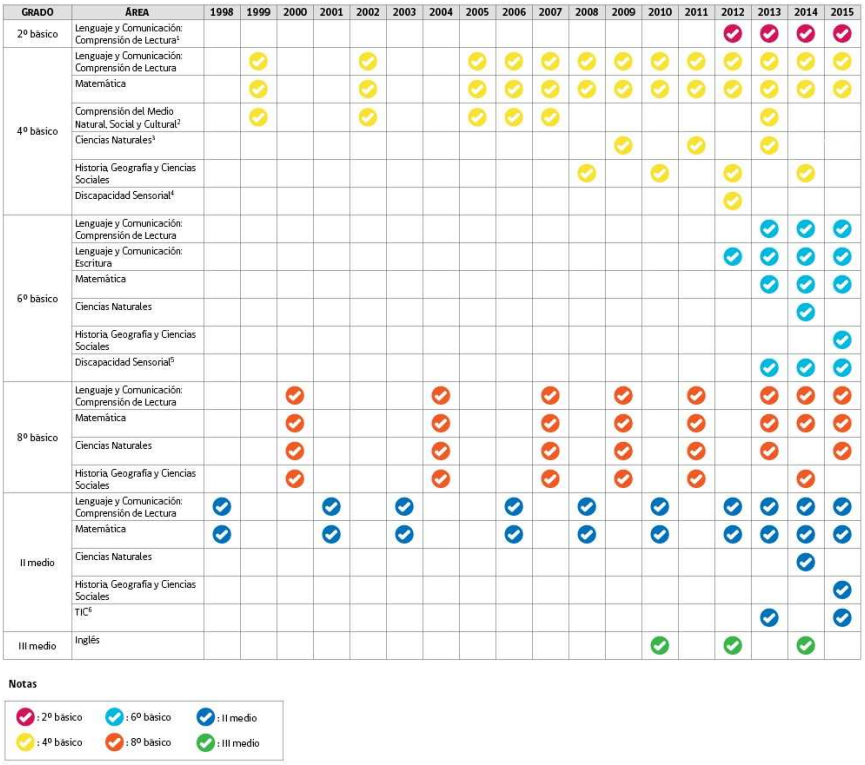
Taut, S., M. V. Santelices, C. Araya, and J. Manzi (2011, December). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. *Studies in Educational Evaluation 37*(4), 218–229.

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: Simce Tests by Year, Grade, and Subject (in Spanish)



GRADO	ÁREA	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2º básico	Lenguaje y Comunicación: Comprensión de Lectura[1]															●	●	●	●
4º básico	Lenguaje y Comunicación: Comprensión de Lectura		●			●			●	●	●	●	●	●	●	●	●	●	●
	Matemática		●			●			●	●	●	●	●	●	●	●	●	●	●
	Comprensión del Medio Natural, Social y Cultural[2]		●			●			●	●	●						●		
	Ciencias Naturales[3]												●		●		●		
	Historia, Geografía y Ciencias Sociales											●		●		●		●	
	Discapacidad Sensorial[4]															●			
6º básico	Lenguaje y Comunicación: Comprensión de Lectura																●	●	●
	Lenguaje y Comunicación: Escritura															●	●	●	●
	Matemática																●	●	●
	Ciencias Naturales																	●	
	Historia, Geografía y Ciencias Sociales																		●
	Discapacidad Sensorial[5]																●	●	●
8º básico	Lenguaje y Comunicación: Comprensión de Lectura			●				●			●		●		●		●	●	●
	Matemática			●				●			●		●		●		●	●	●
	Ciencias Naturales			●				●			●		●		●		●		●
	Historia, Geografía y Ciencias Sociales			●				●			●		●		●			●	
II medio	Lenguaje y Comunicación: Comprensión de Lectura	●			●		●			●		●		●		●	●	●	●
	Matemática	●			●		●			●		●		●		●	●	●	●
	Ciencias Naturales																	●	
	Historia, Geografía y Ciencias Sociales																		●
	TIC[6]																●		●
III medio	Inglés													●		●		●	

Notas

● : 2º básico ● : 6º básico ● : II medio
● : 4º básico ● : 8º básico ● : III medio

[1] El Consejo Nacional de Educación ha decidido realizar la prueba Simce Comprensión de Lectura de 2º básico para el año 2015.

[2] Hasta el año 2007 se evaluaron los conocimientos y habilidades del área Comprensión del Medio Natural, Social y Cultural, señalados en los Objetivos Fundamentales y Contenidos Mínimos Obligatorios correspondientes al Nivel Básico 1 y Nivel Básico 2 del Marco Curricular, según se establece en el Decreto N.º 232 del año 2002.

A partir del año 2008 se aplican de manera intercalada las pruebas referidas a la *Comprensión del Medio Social y Cultural* (nombrada en el calendario como Historia, Geografía y Ciencias Sociales) y *Comprensión del Medio Natural* (nombrada en el calendario como Ciencias Naturales) del área de Comprensión del Medio Natural, Social y Cultural.

[3] El Consejo Nacional de Educación ha aprobado la propuesta del Ministerio de Educación de eliminar la prueba Simce Ciencias Naturales de 4º básico 2015.

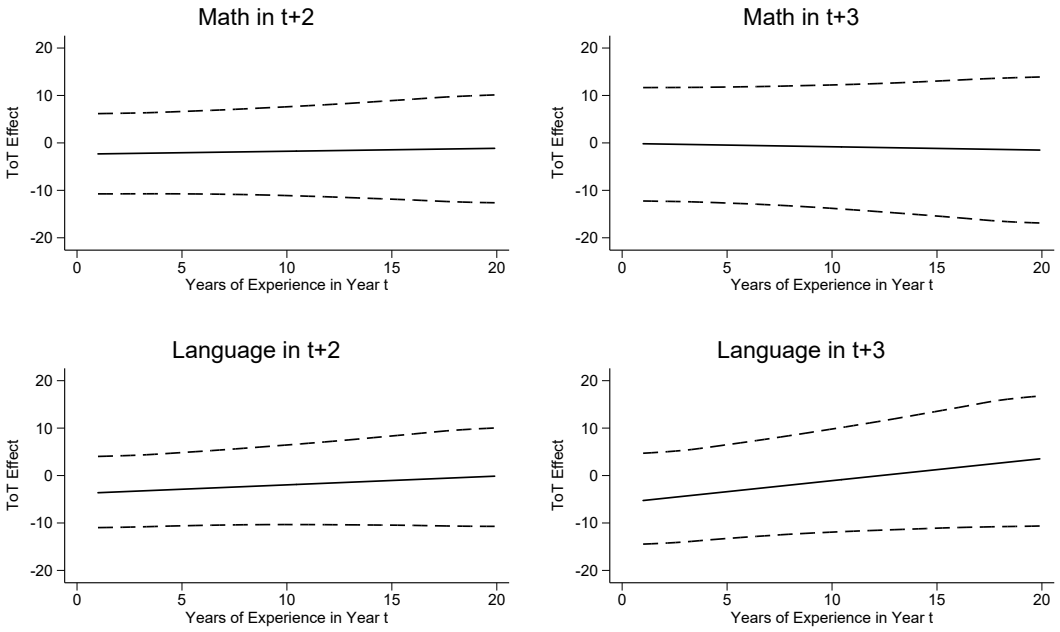
[4] Desde el año 2009 hasta el 2012 se aplicó regularmente la evaluación para estudiantes con discapacidad sensorial en 4º básico en las regiones de Valparaíso, Metropolitana de Santiago y Biobío.

[5] A partir de 2013 se evalúa a todos los alumnos con discapacidad sensorial de 6º básico del país, lo que permitirá ampliar los resultados de aprendizaje existentes hasta 2012 en esta área, y entregar un panorama más acabado de los aprendizajes alcanzados y de su contexto en la modalidad de educación especial en Chile.

[6] Esta prueba está en proceso de revisión para alinear más claramente su contenido con las bases curriculares vigentes. En régimen, esta evaluación será un estudio muestral realizado por Enlaces en II año de Educación Media, aplicado cada 3 años, a partir del año 2016. Esto se traduce en que esta prueba no se aplicará el 2015.

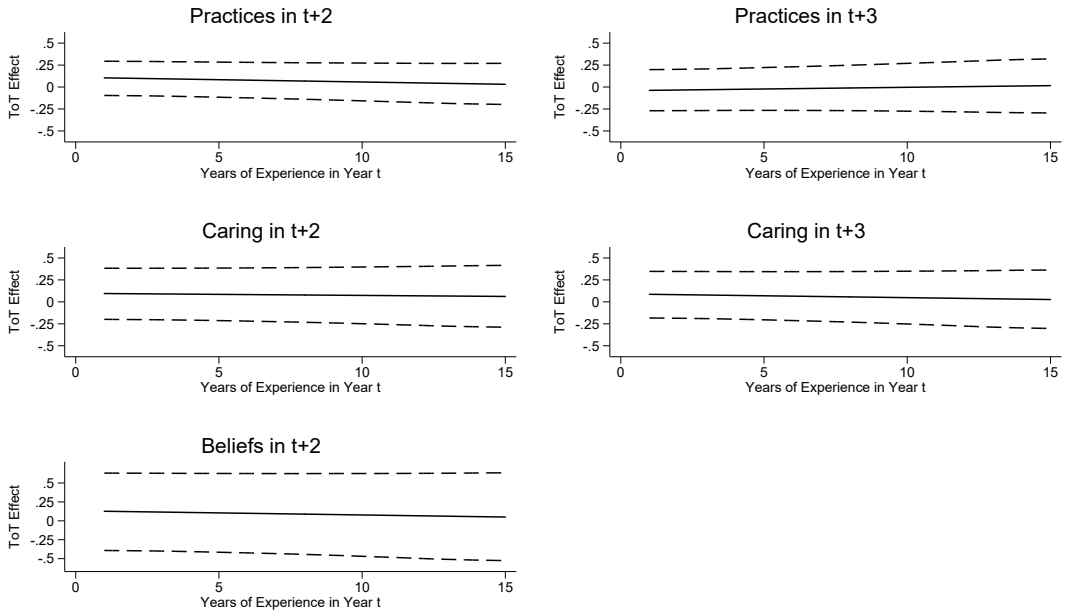
Source: Agencia de Calidad de la Educación (2015)

Figure A2: ToT Effects of Teacher Re-evaluations on Student Learning, by Teaching Experience



Note: Difference-in-Difference estimates of teacher re-evaluation effects (ToT) on student performance in the year of and one year after the re-evaluation, by teachers' work experience. 95% confidence interval shown with dashed lines.

Figure A3: ToT Effects of Teacher Re-evaluations on Mediating Variables, by Teaching Experience



Note: Difference-in-Difference estimates of teacher re-evaluation effects (ToT) on teaching practices (student-reported), teachers' beliefs, and teachers' level of caring (parent-reported) in the year of and one year after the re-evaluation, by teachers' work experience. 95% confidence interval shown with dashed lines.