UNIVERSIDAD TORCUATO DI TELLA

Escuela de Gobierno

ESCUELA DE GOBIERNO

DOCUMENTOS DE TRABAJO 2019/04

# IS THE REMEDY WORSE THAN THE DISEASE?

# THE IMPACT OF TEACHER REMEDIATION ON TEACHER AND STUDENT PERFORMANCE IN CHILE

MARÍA LOMBARDI

SEPTIEMBRE 2019

Documentos de trabajo: https://bit.ly/36RfJJ0

UTDT: Av. Figueroa Alcorta 7350, C1428BCW Buenos Aires, Argentina

# Is the Remedy Worse Than the Disease? The Impact of Teacher Remediation on Teacher and Student Performance in Chile

María Lombardi*
Universidad Torcuato Di Tella

September 2019

**Abstract**

I study the impact of remedial training for low-performing teachers in Chile. Taking advantage of the fact that assignment to remediation is mainly based on teacher evaluation scores, I use a fuzzy regression discontinuity design and find that teachers barely assigned to remediation improve their pedagogical practices as measured by their next evaluation scores. While there is suggestive evidence that these teachers' students obtain higher standardized test scores after the training is complete, this result is not robust, and the suggestive positive impact disappears after one year. I also find that during the year of their teacher's reevaluation, the students of teachers assigned to remedial training obtain significantly lower test scores. Teachers assigned to remediation report lower prestige and job satisfaction, suggesting that the stigma of being labeled as a low performer leads teachers to put more effort into preparing their teaching evaluations, causing a temporary drop in student learning.

*JEL Classifications*: I21, J24, M53
*Keywords:* education, teachers, training, Chile.

*Contact: mlombardi@utdt.edu. School of Government, Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, C1428BCW Buenos Aires, Argentina. I am grateful to Felipe Barrera-Osorio, Andy de Barros, Ana Figueiredo, Niklas Heusch, David Jaume, Gianmarco León, Marco Nieddu, Stephan Litschig, Joe Vecci, and to seminar and workshop participants at CAF, FGV-EPGE, Universitat de Girona, the University of Gothenburg, the ASWEDE Conference, the RIDGE Workshop in Public Economics, the Nordic Summer Institute in Labor Economics, and the RIDGE Workshop in Impact Evaluation. The views expressed herein are those of the author.

# 1 Introduction

A sizable share of teachers in low- and middle-income countries lack the knowledge and pedagogical skills for effective teaching (Bruns and Luque, 2015; Bold et al., 2017). Raising the calibre of teachers is critical, as teacher quality is a key determinant of student learning (Rockoff, 2004; Hanushek and Rivkin, 2010; Araujo et al., 2016). While governments should strive to recruit better teachers, improving the selection of new entrants will not lead to a meaningful increase in the average quality of teachers in the short run, as a large share of public school teachers are civil servants with permanent contracts. Potential policies to upgrade the skills of current teachers include evaluating them (Taylor and Tyler, 2012), giving them feedback (Muralidharan and Sundararaman, 2010; De Hoyos et al., 2017), and providing them with training. Even though in-service training for teachers is a major component of non-salary expenditures in most educational systems (Bruns and Luque, 2015),[1] many of these programs go unevaluated, and the evidence from those that are assessed is often discouraging (Fryer, 2017; Popova et al., 2018). A potential pathway to improve the efficiency of in-service training is to use the results of teacher evaluations to target the teachers most in need of training. Training programs can then be tailored to address the weaknesses of each teacher. However, there is no evidence on whether remedial training can improve the skills of low-performing teachers.

This paper investigates the impact of remedial training for low-performing teachers in Chile on their subsequent performance in teaching evaluations and their students' achievement in standardized tests. In 2003, the Chilean Ministry of Education introduced evaluations for public school teachers. Teachers are periodically evaluated using several instruments such as classroom observations, an extensive lesson plan prepared by the teacher, and peer and supervisor assessments.[2] Teachers with a weak performance in this evaluation (roughly the bottom tercile) are required to

[1] In the 2013 OECD Teaching and Learning International Survey of 38 developed and developing countries, 88% of teachers in lower secondary education report engaging in some professional development in the last year (OECD, 2014).

[2] Several studies examining the validity of Chile's teacher evaluation instruments find that teachers' scores are correlated with value added in terms of students' standardized test scores (Alvarado et al., 2012; Taut et al., 2016; Bruns et al., 2016).

attend remedial training until their next evaluation in four years' time. Taking advantage of the fact that assignment to remediation is primarily based on a cutoff rule, I use a fuzzy regression discontinuity (RD) design to obtain causal estimates of the impact of remedial training for teachers first evaluated in 2004-2010. Since not all teachers who were assigned to remedial training take part in it, my estimates capture the impact of being assigned to remediation.[3]

Providing teachers with remedial training could encourage them to improve their skill set, but teachers' pedagogical skills will not improve if these training programs are ineffective. In contrast to the standard in-service training that teachers in most educational systems receive, remediation requires identifying a group of teachers whose performance is considered inadequate. Singling teachers out in this way has the potential of creating social stigma (Koedel et al., 2017), which could incentivize teachers to exert more effort, or discourage them and worsen their performance. Using individual-level data on the 38,000 teachers first evaluated between 2004 and 2010, I find that those who are barely assigned to remedial training after their first appraisal have higher reevaluation scores than those who barely avoided training. Importantly, I show that there is no differential attrition or sorting that could bias my estimates, and provide evidence that the density of teachers' scores in the first evaluation (i.e., the running variable) is smooth across the cutoff, as are the observable characteristics of teachers and the schools they work for. These results are robust to the inclusion of these baseline covariates, and to the choice of bandwidth and functional form. I also find that teachers assigned to remedial training are significantly less likely to report increased prestige and job satisfaction as a consequence of their first evaluation results, indicating the existence of stigma associated with remediation. It is important to mention that low-performing teachers are not only assigned to remediation, they are also excluded from applying for temporary wage increases. Although I cannot disentangle the impact of these two treatments, I present evidence that suggests that the improvement in evaluation scores is, at best, only partially driven by differential salary increases.

[3]Survey responses show that approximately 20% of teachers who were assigned to remediation did not participate in it. However, this information is self-reported, and is only available for a subsample of teachers. I thus estimate the intent-to-treat effect of remediation.

While remedial training improves the performance of teachers in their evaluations, it is unclear whether this translates into better teaching. To assess the causal impact of teacher remediation on student learning, I use individual data on student performance in Chile's yearly nationwide standardized tests in 2005-2016, and compare the test scores of students whose teacher was barely assigned to remediation with those whose teacher barely avoided it.[4] Although I find suggestive evidence that the students whose teachers were barely eligible for remedial training obtain higher test scores the year after their teacher completes remediation, these results are not robust. Furthermore, there is no statistically significant difference in test scores the following year. I also find that on the year of their teachers' reevaluation (also the fourth and final year of training), the students of teachers assigned to remediation score around 0.12 standard deviations lower on their standardized tests. Participating in remedial training and preparing for the teaching evaluations take time, and teachers' workload is not reduced to compensate for this (Taut et al., 2011; Taut and Sun, 2014). Such a sizable decrease in student learning is thus consistent with a story in which teachers assigned to remediation face the stigma of being labeled as low performers, making them more likely to allocate time away from their duties to prepare for their reevaluation. Taken together, these results indicate that even though remedial training can have a modest impact on the quality of teaching, it can also have unintended consequences that counteract the program's success.

Although there is an extensive literature on the topic of in-service-training for teachers,[5] there are no studies examining the impact of remedial training for low-performing teachers. The closest paper is Jacob and Lefgren (2004), who examine the impact of a reform that placed a subset of elementary schools in Chicago with low reading scores on academic probation. Probation schools received special funding for staff development, as well as technical assistance and enhanced monitoring. Since assignment to probation was based on a cutoff rule, the authors also

[4] Importantly, standardized test scores are not an input of teacher evaluations, and thus provide a clean measure of student achievement.

[5] The meta-analyses of teacher training programs in high-income countries by Yoon et al. (2007) and Fryer (2017) find mixed results, and show that the programs that were most beneficial where those that provided the most hours of training, and had precise training and curriculum materials for implementers to follow. A meta-analysis of 33 programs in low- and middle-income countries by Popova et al. (2018) shows that the in-service training programs that increase student learning usually link participation to career incentives, focus on subject-specific pedagogy, include face-to-face training, and incorporate lesson enactment where teachers can practice with one another.

employ a fuzzy RD design, and find that teacher training had no impact on students' reading or math scores. However, the remedial training in Jacob and Lefgren (2004) was granted to *schools* with low standardized test scores, whereas training in Chile is provided to *teachers* with low scores in an evaluation of their pedagogical skills. The metric used to assign teachers to remediation (school-level standardized test scores vs. individual evaluations of teachers' pedagogical skills) and the level at which remediation is provided (every teacher in the school vs. a subset of low-performing teachers) could matter for how teachers respond to remedial training, and how this impacts their students. To the best of my knowledge, this is the first study analyzing the impact of training specifically targeting low-performing teachers.

The paper is organized as follows. Section 2 describes the Chilean teacher evaluation system and its consequences, and Section 3 outlines the regression discontinuity design used for estimating the causal impact of remedial training. Section 4 describes the data and provides details on the sample and its characteristics, and Section 5 presents the main results. Section 6 presents several validity and robustness checks, and Section 7 discusses the interpretation of the findings. Section 8 concludes.

# 2 Teacher Evaluations and Remedial Training in Chile

## 2.1 Teacher Evaluations

In 2003, the Chilean Ministry of Education introduced teaching evaluations in public schools, which employ around 40% of the country's teachers. The evaluation system was designed around the Ministry's previously developed national framework which defined the standards for the teaching profession (Santiago et al., 2013). Teachers are evaluated every four years during the second half of the school year, using four instruments aimed at capturing different aspects of their performance: (i) a teaching portfolio, (ii) a peer assessment, (iii) a self assessment, and (iv) a supervisor assessment.

The teaching portfolio, which accounts for 60% of a teachers' evaluation score, is composed

of two separate parts. The first is a lesson that is videotaped by a cameraperson sent by the Ministry on an pre-appointed day. The second part of the portfolio is composed of different elements that make up an eight-hour learning unit that teachers have to plan and implement, in the grade and subject they conduct most of their teaching. Teachers must present a written lesson plan for each of the classes in this learning unit, an assessment of student learning, and must answer several questions reflecting on their pedagogical decisions and on their students' assessment results. Portfolios are then blindly graded by trained teachers with at least five years of classroom experience in the same subject area and grade level as the evaluated teacher.[6] The second element of the evaluation, accounting for 20% of the final score, is a peer interview conducted by a trained teacher from a different school who instructs the same subject and grade.[7] A self-assessment questionnaire accounts for 10%, and the remaining 10% of the score is made up of a supervisor assessment completed by the school principal and the school's head of the relevant technical-pedagogical unit. A few studies examining the validity of these instruments show that teachers' portfolio scores have the strongest association with value added in terms of students' standardized test scores (Alvarado et al., 2012; Taut et al., 2016; Bruns et al., 2016).[8] While supervisor assessments are also correlated with value added, the ratings in the assessments conducted by a peer or the teacher himself are poor predictors of student test score gains.

Teachers' final evaluation scores, ranging from 1 to 4, are a weighted average of these four instruments. Teachers are then divided into four categories (Unsatisfactory, Basic, Competent and Outstanding) using fixed cutoffs. According to the Ministry of Education's guidelines (CPEIP,

[6]Local universities are in charge of training the evaluators and running the portfolio grading process, which follows strict rubrics stemming from the Ministry's Framework for Good Teaching. The videotaped lessons are evaluated on three dimensions: classroom learning environment, lesson structure, and quality of interactions. Raters must evaluate the written lesson plan in terms of four dimensions: unit organization, analysis of the unit's lessons, quality of the assessment, and reflection upon the assessment results. The videotaped lesson and written lesson plan are graded by separate raters, and an overall portfolio score is calculated by averaging these seven dimensions. Further details on the portfolio grading process can be found in Taut and Sun (2014).

[7]Every year, teachers must apply to become a peer evaluators, and those who are selected are trained on how to perform this assessment. The appraisal conducted by these teachers is a scripted interview in the evaluated teacher's school, and lasts for approximately one hour (Manzi et al., 2011).

[8]Several high quality studies have also found a positive relationship between teacher quality measures extracted from classroom observation and value added (for example, Kane et al. (2011) and Kane et al. (2013) in the US, and Araujo et al. (2016) in Ecuador). It should be noted, however, that classroom observation rubrics are poor predictors of student performance in terms of complex cognitive skills or social-emotional competencies (Kraft, 2019).

2008), teachers with a Competent performance display an adequate professional achievement that complies with the minimum requirements for exercising their teaching duties. Figure 1 shows the distribution of evaluation scores and the corresponding categories for all primary and secondary school teachers evaluated between 2004 and 2010.[9] In this period, 34% of teachers had a performance below the minimum required level (in only 2% of cases the score was Unsatisfactory, and in the remaining 32% it was Basic). The majority of teachers (58%) obtained a score which entitled them to a Competent rating, and less than 8% received an Outstanding score. During the summer break, an evaluation committee composed of peer evaluators and the municipal education authorities convenes in every municipality to analyze the results of the evaluation process, and decides whether to ratify or modify them based on contextual considerations.[10] Teachers' final categories are modified in approximately 5% of cases, and most of these changes imply bumping teachers up to the next category. Importantly, these committees only change a teacher's final category, not the underlying numeric score. At the start of the following school year, teachers receive a detailed report with the result of their evaluation (i.e., their category), and feedback on their performance in each of the evaluation instruments as compared to the minimum required standard, with a particular emphasis on the portfolio.[11] Teachers also learn whether their final score was ratified or modified by their municipal evaluation committee, but they do not find out the numeric score they obtained.

The rollout of teacher evaluations was gradual, as displayed in Appendix Figure A.3, reaching all municipalities, grade levels and subjects in primary and secondary school by 2009.[12] However, coverage is not universal. First-year teachers are not evaluated, and those who are three or less years away from retiring can opt out. Both of these groups constitute around 16% of public

[9] Appendix Figure A.1 depicts the distribution of scores in each of the evaluation instruments for the same sample of teachers.

[10] Public education in Chile is administered by municipal authorities, and there are 346 municipalities.

[11] Appendix Figure A.2 shows a translated example of the feedback that teachers get on one of the dimensions in which their portfolio is evaluated. School principals and the heads of the municipal school board also receive a report summarizing the results of the evaluation for all teachers who were evaluated the year before in that school or municipality, respectively. An example of the report that teachers and school principals receive (in Spanish) can be found here and here, respectively.

[12] Following the enactment of Law 19,961 in 2004 and Decree 192 in 2005, evaluations became mandatory. In 2003-2004, participation was voluntary, and only a few municipalities participated.

school teachers in a given year. Furthermore, teachers can postpone their assessment if they act as peer evaluators, for personal reasons, or by simply refusing to be evaluated. Although 11% of teachers up for evaluation in 2006 refused to be tested, these numbers quickly went down because teachers who decline to participate are automatically granted an Unsatisfactory rating, and those with three consecutive Unsatisfactory ratings are dismissed (Manzi et al., 2011).

## 2.2 Remedial Training and Other Consequences of Teacher Evaluations

Teaching evaluations can be of a formative or a summative nature. Formative evaluations have low stakes, and their main purpose is to provide diagnosis, feedback and training to improve the quality of teaching. Summative assessments, on the contrary, have higher stakes attached to them, and aim to punish low performers and reward high achievers. Although the Chilean teacher evaluation system presents features of both types of assessments, until 2011 it emphasized the formative aspects, which targeted teachers with an Unsatisfactory and Basic performance. These teachers were required to attend annual remedial training courses until the completion of the next evaluation in four years' time.[13]

The remedial training courses are designed and implemented by municipal school authorities, but are funded by the Ministry of Education. Every year, the municipal school authorities must present a proposal for their remediation activities, and the Ministry of Education dispenses the funds only after reviewing and approving these plans. Municipalities receive approximately 127 and 400 dollars per year (in dollars of 2018) for every teacher with a Basic and Unsatisfactory rating, respectively.[14] Since teachers with a Basic rating undergo four years of training, this amounts to 508 dollars per teacher. The total funds devoted to training a teacher with a Basic rating represent, on average, 14% and 25% of the annual tuition fee for a degree in education in a university or professional institute, respectively.[15] The take-up of remedial training by munici-

[13]Teachers with an Unsatisfactory score are the exception. Since they were reevaluated after just one year, they only received one year of remediation.

[14]These are averages over 2005-2014, the years in which the teachers in my sample attended remedial training. The nominal cost per teacher increased over time, but it was relatively constant in real terms.

[15]The annual tuition fees for every degree offered by every university and professional institute in Chile in 2018

pal authorities presented some delays. In 2005, for example, 72% of municipalities with teachers evaluated at the Basic or Unsatisfactory level implemented these training programs. By 2010, the compliance rate rose to 90%.[16]

Municipal school authorities have considerable freedom in terms of the format and content of their remediation activities. In an annual survey conducted during the evaluation process, teachers who obtained an Unsatisfactory or Basic rating in their previous evaluation are asked about the characteristics of the remediation activities they participated in. The results of this survey are summarized in Appendix Table A.1. As the questions in this survey changed over time, these statistics only apply to teachers who were reevaluated in the years indicated in the last column. Importantly, response rates are quite high, as shown in the third column. Most teachers responded that the remediation activities covered lesson planning (71%), student assessment (67%), and reflection about their own teaching practices (53%). A smaller fraction responded that remediation was aimed at improving the learning environment (35%), enhancing pedagogical skills (28%), or mastering content related to the subject teachers specialize in (20%). With regards to the type of activities, 67% of teachers responded that remediation consisted of lectures, and a smaller proportion (19%) participated in group discussions. A very small share of teachers participated in other forms of remediation such as mentoring or coaching, role-play or simulation, or analysis of teaching practices in videotaped lessons.

Until 2011, the main summative consequences of teacher evaluations applied to those with an Unsatisfactory score. These teachers were reevaluated after just one year and would be terminated if they got three consecutive Unsatisfactory scores. Teachers who had a Competent or Outstanding rating were eligible for a temporary salary increase that lasted until the next evaluation ("Asignación Variable al Desempeño Individual", henceforth AVDI) and was granted to those who applied for it and passed a content mastery test. Almost half of the teachers who were eval-

can be found in http://portal.beneficiosestudiantiles.cl/aranceles-de-referencia (last accessed December 8, 2018).

[16]Given that non-complying municipalities were typically smaller, most teachers evaluated with an Unsatisfactory or Basic rating worked in a district that offered remediation. In 2005, for instance, 85% of teachers assigned to remediation taught in a municipality where these training programs were implemented, and by 2010 this percentage rose to 96%.

uated for the first time in 2004-2010 with a Competent or Outstanding rating received this salary increase, which was 3.6% on average.[17] In 2011, the teacher evaluation system was reformed, and harder accountability measures were introduced. In particular, the threat of dismissal for teachers with an Unsatisfactory performance became stronger, and some punitive consequences were introduced for teachers obtaining a Basic score.[18] I therefore focus my analysis on teachers evaluated for the first time in 2004-2010, when the consequences of the evaluation process were mostly formative. Figure 2 summarizes the evaluation process and its consequences for teachers evaluated during this time; it focuses on those who obtained a Basic or Competent score because my empirical strategy is concentrated on these two groups of teachers (around 90% of those first evaluated in 2004-2010), as explained in further detail in Section 3.

# 3 Estimation Strategy

I estimate the causal impact of being assigned to teacher remedial training by comparing teachers who received a Basic rating in their first evaluation with teachers who received a Competent rating instead, as the former were assigned to remedial training but the latter were not. Taking advantage of the cutoff rule assigning teachers to the Basic or Competent category, I employ a regression discontinuity design, comparing the subsequent performance of teachers just above and just below the relevant cutoff score. As can be seen in Figure 3, crossing the Basic/Competent threshold leads to an 82 percentage point drop in the probability of obtaining a Basic score and thus being assigned to remediation. Although this discontinuity is large, the probability of obtaining a Basic rating does not jump from 1 to 0 because 15% of teachers with scores right below the cutoff are bumped up to the Competent category,[19] and 3% of teachers with scores to the right of the cutoff

[17] Around 65% of the these teachers applied for the AVDI, and 78% of those who applied passed the minimum threshold for receiving the AVDI award. More than 60% of those who passed had a 2% salary increase, and the remaining 39% and 1% obtained a 6% and 10% raise, respectively.

[18] Starting 2011, teachers who obtain two consecutive Unsatisfactory ratings or three consecutive ratings below Competent are dismissed. Moreover, teachers with a Basic rating are reevaluated after only two years. These measures are not retroactive, meaning they only applied to the evaluations conducted after 2011.

[19] Controlling for the overall score, teachers who are bumped up to the Competent category (i.e., the never takers) are significantly more likely to have refused being evaluated in the past (12 percentage points over a mean of around

get demoted to the Basic category. I thus use a fuzzy RD specification, and instrument assignment to remediation using the following first-stage regression:

$$Basic_{i,t_0} = \alpha_0 + \alpha_1 I\left\{S_{i,t_0} - c \geq 0\right\} + \alpha_2 f(S_{i,t_0} - c) + \alpha_3 I\left\{S_{i,t_0} - c \geq 0\right\} \times f(S_{i,t_0} - c) + U_{i,t_0}, \quad (1)$$

where $Basic_{i,t_0}$ is a dummy variable which takes the value of 1 if teacher *i* obtained a Basic rating in his/her first evaluation in year $t_0$ (after revision by the municipal committee) and 0 otherwise, with $t_0$ ranging from 2004 to 2010. $I\left\{S_{i,t_0} - c \geq 0\right\}$ is an indicator variable for whether the teacher's first evaluation score ($S_{i,t_0}$) was above the Basic/Competent threshold *c*, and $f(S_{i,t_0} - c)$ is a polynomial function of this score centered around the cutoff. I include this polynomial by itself and interacted with the indicator variable to allow the relation between the outcome and the running variable to vary at both sides of the cutoff. I first examine the impact of obtaining a Basic rating on the performance of teachers in their second evaluation by estimating the following equation:

$$E_{i,t_0+x} = \beta_0 + \beta_1 \widehat{Basic}_{i,t_0} + \beta_2 f(S_{i,t_0} - c) + \beta_3 I\left\{S_{i,t_0} - c \geq 0\right\} \times f(S_{i,t_0} - c) + \epsilon_{i,t_0+x}, \quad (2)$$

where $\widehat{Basic}_{i,t_0}$ is predicted using Equation (1), and $E_{i,t_0+x}$ is an outcome from teacher *i*'s second evaluation in year $t_0 + x$. I can only perform this analysis on teachers who were reevaluated (65% of the teachers first evaluated between 2004 and 2010). I discuss the causes of attrition in Section 6.1, and show that the probability of being reevaluated is smooth across the cutoff. Even though teachers are supposed to be reevaluated after four years (i.e., *x* should be equal to 4), some teachers take the assessment after five (18%) or more (7%) years. The outcome variables are the teacher's final numeric score, an indicator for whether the teacher's score was above the Basic/Competent cutoff, and the score in each of the four components (portfolio, peer assessment, self assessment

3%), more likely to teach art or music as compared to other subjects (14 percentage points), and less likely to teach in lower primary school (2 percentage points). Bumped-up teachers have higher portfolio grades (0.33 points out of 4), and a lower score in the self and peer assessments (0.06 and 0.09 points out of 4, respectively). There are no other observable characteristics of teachers or the schools they work for that are correlated with the likelihood of being bumped up.

and supervisor assessment). Although remedial training is meant to be mandatory for individuals who obtain a Basic rating, participation is high but not universal. As shown in Appendix Table A.1, 79% of the teachers with a Basic score that were reevaluated after 2009 reported that they attended remedial training courses. Since information on attendance is self reported and only available for a subsample of teachers,[20] the treatment variable in these regressions is an indicator for being *assigned* to remedial training, and so $\beta_1$ measures the intent-to-treat effect of remediation.[21] A potential problem for uncovering the intent-to-treat impact of remediation, however, is that individuals with a Basic score are not only eligible for remedial training, they are also excluded from applying for a temporary salary increase. Therefore, $\beta_1$ captures the effect of being eligible for remediation and ineligible for a salary increase. In Section 7.2 I attempt to unbundle the impact of these two treatments, and provide suggestive evidence that the results are, at best, only partially driven by differential salary increases.

After identifying the effect of remedial training on teachers' reevaluation results, I use the same empirical strategy to study the impact on the standardized test scores of these teachers' students. Using individual student data and following the previous estimation strategy, I run the following regression:

$$Y_{j,i,g,m,t_0+y} = \gamma_0 + \gamma_1 Basic_{i,t_0} + \gamma_2 f(S_{i,t_0} - c) + \gamma_3 I\{S_{i,t_0} - c \geq 0\} \times f(S_{i,t_0} - c) + \zeta_{j,i,g,m,t_0+y}, \quad (3)$$

where $Y_{j,i,g,m,t_0+y}$ is the score that student *j* with teacher *i* in grade *g* got in a standardized test on subject *m*, taken *y* years after the teacher's first evaluation. Given that the impact of training likely depends on the time passed since a teacher started remediation, I estimate separate regressions for each of the six years after teachers' first evaluation. I express test scores as z-scores, standardizing by subject, grade and year, so that $\gamma_1$ captures the standard deviation change in test scores

[20] The question on attendance to remedial training was only included in 2010, excluding 20% of teachers with a Basic rating who were reevaluated in 2008-2009. Furthermore, 9% of teachers did not respond this question.

[21] Almost all the non-compliance with the assignment rule comes from teachers being bumped up (i.e., never takers), and so $\hat{\beta}_1$ is close to capturing the average treatment effect on the treated at the cutoff, where the treatment involves being assigned to remediation.

associated with having a teacher assigned to remedial training. I run these regressions pooling all subjects and grades, but also run separate regressions for math and language in just $4^{\text{th}}$ grade, the only that is tested every year.

I use a local linear regression in my preferred specification, although I also present estimations using a quadratic polynomial of the running variable as a robustness check. Following Calonico et al. (2014), I run these regressions over the MSE-optimal bandwidth using a triangular kernel, and present bias-corrected coefficients and robust bias-corrected standard errors. Since assignment to remedial training is done at the municipality level, I cluster my standard errors by the teacher's municipality in Equation (2). When I run my analysis using student test scores as the outcome variable, I cluster the standard errors by the student's school instead. For robustness, I also present the results of regressions including evaluation-year fixed effects and teacher and school characteristics measured at the year of the teacher's first evaluation.

## 3.1 Internal Validity

In any RD estimation, the basic identifying assumption is that the unobservable determinants of teacher and student performance vary smoothly as a function of teachers' first evaluation scores, the running variable (Lee and Lemieux, 2010). For this assumption to hold, there must be no systematic manipulation of the running variable around the threshold, requiring in this context that teachers and raters have imprecise control over teachers' overall evaluation scores. Despite the fact that teachers and those who rate them can influence a teacher's score, it is highly unlikely that they can anticipate whether the teacher will end up just below (or above) the threshold, and then modify the part of the score that they control to ensure that the overall score crosses this threshold. The main reason for this is that teachers, peer evaluators, supervisors and portfolio raters do not have access to the scores of any of the other evaluation components when they are making their assessment. The last assessment instrument to be graded is the portfolio, and both of its components are rated by different people who do not know the identity of teachers or the scores they obtained in the rest of the evaluation.

To test the validity of this assumption, I start by visually inspecting the histogram of the running variable to see if there is any bunching around the threshold. As displayed in Panel A of Figure 4, the distribution seems to be smooth around the Basic/Competent cutoff. I formally test whether the density of the running variable is continuous around the neighborhood of the Basic/Competent cutoff using the approach of McCrary (2008), and do not find a statistically significant jump in the density at this threshold.[22] To further confirm that there is no systematic manipulation, I also test whether a host of pretreatment characteristics of teachers and the schools they work for are continuous around this cutoff, using the same specification as in Equation (2). Table 1 shows the results of fuzzy RD regressions testing whether each of these characteristics are balanced for teachers with a Basic and Competent rating in a neighborhood around the cutoff. Additionally, Appendix Figures A.4-A.10 plot the relation between these characteristics and the running variable. Although teacher covariates are unavailable for 4% of teachers and school covariates are missing for 0.3%, the likelihood of missing this information is continuous at the cutoff between a Basic and Competent rating. As can be seen in Table 1, only 1 out of 27 covariates is imbalanced at conventional significance levels. In particular, teachers with a Basic rating are close to 5 percentage points more likely to come from a school that won the previous edition of SNED, a nationwide teacher pay-for-performance program. Importantly, there are no other baseline differences in the characteristics of the schools teachers work for, such as the number of students, the teacher-student ratio, students' average socioeconomic status, standardized test scores, or whether the school is located in an urban setting. For robustness, I also report the results of regressions including evaluation-year fixed effects and these baseline covariates.[23]

While there are no systematic differences in the baseline characteristics of teachers barely assigned and not assigned to remediation, my empirical analysis is conducted using outcomes measured several years after teachers are first evaluated. Causal interpretation of my estimates also

[22]The point estimate for the difference in log heights at the threshold is -0.006 (s.e. 0.021). The graphical depiction of the McCrary test is shown in Panel B of Figure 4.

[23]The only covariate not included in the regressions with controls is the average standardized test scores of the school the teacher worked for on the year of the first evaluation. I do not include this variable because it is not reported for schools with few students, and is thus missing for almost 7% of teachers. In the regressions using student test score data, I also control for the student's gender, and the average socioeconomic status of the students in the class.

requires that teachers do not exhibit differential attrition or sorting. I discuss this issue in detail in Section 6.1, and show that there is no differential attrition or sorting of teachers into different jobs, schools or classes across the cutoff.

# 4 Data and Descriptive Statistics

I rely on two sets of data provided by the Chilean Ministry of Education to estimate the causal impact of remedial training on teachers' performance in their second evaluation. First, a database on teachers evaluated since 2004 with information on the grade level and subject in which teachers were evaluated, their overall score (both the numeric score and the final category), their score in each of the evaluation instruments, and responses from a teacher survey conducted during the evaluation process.[24] I start out with 58,585 primary and secondary school teachers who were first assessed in 2004-2010, and restrict my sample to the 37,866 teachers who were reevaluated before 2016.[25] Secondly, I obtain teacher covariates from a database with every job held by teachers in the Chilean school system. This database contains a myriad of teacher characteristics, such as their age, gender, whether they hold a degree, and their years of experience. It also contains information on the features of the position, such as the weekly number of contractual hours, the type of contract (i.e., if the teacher is a civil servant with a permanent contract or a contract teacher), if the teacher also holds an administrative position, and unique school identifiers.[26] Both databases have unique teacher identifiers that allow tracking teachers across years and permits the merging of different datasets.

To study the impact of remedial training on student learning, I rely on student-level data

[24] Chile's teacher evaluations started in 2003, and covered lower primary school teachers ($1^{st}$ to $4^{st}$ grade) in 63 municipalities in that year. Unfortunately, the Ministry of Education does not provide data with the results of the assessments for that year.

[25] Given that there were some changes to the portfolio elaboration and grading process in 2016, I only consider teachers who were reevaluated by 2015.

[26] I also use data from other public databases provided by the Ministry to back out school characteristics, such as whether the school won the nationwide pay-for-performance program, and whether it is located in an urban or rural area. If a teacher works in more than one public school, I focus on the one in which most of his/her teaching is concentrated.

from Chile's nationwide standardized tests ("Sistema de Medición de la Calidad de la Educación", henceforth SIMCE) held in 2004-2016, and merge this information with the database of evaluated teachers using classroom identifiers. Since I am interested in studying the impact on test scores before, during, and after remediation, I look at the performance of the evaluated teachers' students one to six years after the teachers' first appraisal. To make the estimation more tractable, I further limit my sample of teachers to those who were reevaluated after four years. Given that not all grades and subjects participate in SIMCE, my sample only contains $4^{\text{th}}$, $8^{\text{th}}$ and $10^{\text{th}}$ graders, and provides information on their scores in math, language, natural science and social science for $4^{\text{th}}$ and $8^{\text{th}}$, and just math and language for $10^{\text{th}}$ graders.[27]

After excluding teachers who do not teach one of the tested subjects (18%) or grades (17%) in the six year period after their first evaluation, my sample contains 17,953 teachers. It is important to note, however, that almost none of these teachers instructs a tested subject and grade the entire six year period after their first evaluation. In fact, the median teacher instructs a tested grade and subject only two out of the six years. One reason for this is that in $8^{\text{th}}$ and $10^{\text{th}}$ grade, SIMCE is only carried out every other year for most of the analysis period. But even in $4^{\text{th}}$ grade, where students are tested on a yearly basis, most teachers only appear in my sample between one and three years after their first evaluation. The main reason is that Chilean teachers usually rotate across grades, instead of teaching the same grade every year. A limitation of my analysis is that as I run separate regressions for each of the six years after teachers' first evaluation, each of these regressions has a different sample of teachers. I discuss the implications of this for the interpretation of my results in Section 6.1.

Table 2 presents baseline characteristics for: (i) the full sample of primary and secondary school teachers that were evaluated for the first time in 2004-2010; (ii) the subsample of (i) that also took the second evaluation; and (iii) the subsample of (ii) that was reevaluated after four years and taught a grade and subject that participated in SIMCE at some point in the six years after their

[27] Appendix Table A.2 depicts the grades and subjects covered by SIMCE in my period of analysis. I do not consider $2^{\text{nd}}$ and $6^{\text{th}}$ grade because the test was only introduced in these grades at the very end of my analysis period, and only look at math and language for $10^{\text{th}}$ grade students, since there is more than one teacher instructing natural and social sciences in this grade.

first evaluation. The number of observations changes slightly depending on the covariate under consideration, as 4% of teachers do not appear in the database with teacher characteristics and 0.3% do not have school identifiers. Almost 70% of teachers are female, and the average teacher in the full sample is 46 years old, and has 18 years of experience by the time of the first evaluation. Almost all teachers possess a degree, and 69% are civil servants (as opposed to contract teachers). The average working week in the school in which teachers are evaluated is 34 hours long, although 15% of teachers work for more than one school. With regards to the results of their first evaluation, the average score is 2.60 (in a range from 1 to 4). Scores in the self assessment are very high (3.79 on average), followed by the supervisor assessment (3.08) and the peer assessment (2.88). Teachers' scores in the portfolio are considerably lower (2.23). Almost two thirds of these teachers take the second evaluation by 2015, and those who do so are tested after 4.29 years on average.

As shown in column 4 of Table 2, the sample of teachers who are reevaluated is slightly younger and less experienced than the full sample, but there are no other striking differences in terms of their baseline characteristics, their evaluation results, or the characteristics of the schools they worked for. As compared to the sample of reevaluated teachers, those who also teach a grade and subject that was tested in SIMCE (column 7) are less likely to teach in secondary school, and more likely to teach in lower primary school (grades 1-4). This stems from the fact that $4^{\text{th}}$ graders are tested more frequently than those in $8^{\text{th}}$ and $10^{\text{th}}$ grade. They are also more likely to work in rural and thus smaller schools, but do not differ substantially in terms of other characteristics, or in aspects related to the results of their first evaluation. On average, teachers in this restricted sample instruct a SIMCE grade and subject two of the six years after their first evaluation.

# 5 Results

## 5.1 Teachers' Reevaluation Scores

The fuzzy RD estimates displayed in Table 3 show that teachers who were eligible for remedial training got an overall score 0.026 points higher in their second evaluation (out of 4). This discon-

tinuity in teachers' evaluation scores across the cutoff is clearly depicted in Figure 5.[28] The results in column 2 show that teachers assigned to remediation are also 5 percentage points more likely to have a score that crosses the Basic/Competent cutoff, a 7% improvement over the mean. Both of these coefficients are statistically significant at the 5% level, and robust to the inclusion of year fixed effects and teacher and school controls (Panel B), undermining any concerns about manipulation in the running variable.[29] The number of teachers in Table 3 changes across regressions due to variations in the MSE-optimal bandwidth. Importantly, as shown in Section 6.2, the results are robust to using alternative bandwidths.

Breaking down the results reveals that the jump in evaluation scores across the cutoff is completely driven by the portfolio, the evaluation instrument that has the largest weight (60%), is graded in the most objective manner, and has the highest correlation with value-added measures of student achievement (Alvarado et al., 2012; Taut et al., 2016; Bruns et al., 2016). Although the impact on portfolio scores is small in terms of the mean (2% increase), it represents 16% of a standard deviation, and is equivalent to moving from the $50^{\text{th}}$ to the $56^{\text{th}}$ percentile in the distribution of portfolio scores. Further breaking down these results shows that the improvement in portfolio scores is entirely concentrated in the portion devoted to planning and implementing a pedagogical unit, as displayed in Appendix Table A.3. Teachers assigned to remedial training have higher scores in organizing the learning unit, analyzing their pedagogical choices, and reflecting upon their students' assessment results; these impacts represent a 2.6%, 3.2%, and 5.7% increase over the mean, respectively. These findings are remarkably consistent with the survey responses of reevaluated teachers regarding the content of their remedial training courses. As shown in Appendix Table A.1, most teachers responded that the main focus of their remediation activities was lesson planning, student assessment and reflection about teaching.

[28] A potential concern is that these effects are driven by a lower effort from teachers who barely obtained a Competent rating in their first assessment. However, teachers at both sides of the cutoff obtain higher scores in their second evaluation, on average, as the threshold score for obtaining a Competent rating is 2.5. The improvement is simply larger for teachers barely assigned to remediation. Further mitigating this concern, I obtain quantitatively similar results when using a donut hole RD estimation excluding teachers within 0.01 or 0.02 points from the cutoff (results upon request).

[29] The sample used for the estimations in Panel B is slightly smaller (36,199 teachers instead of 37,866) because the information on teacher covariates is missing for 4% of teachers.

## 5.2 Students' Standardized Test Scores

Although teachers obtain higher evaluation scores after being assigned to remediation, it is unclear whether this translates into greater student learning. On the one hand, remedial training may have enhanced teachers' skill set, potentially leading to higher student learning. But if the skills acquired during remediation are only useful for their evaluations (i.e., "training for the test"), there will be no impact on student achievement. Alternatively, training may have had no impact on teachers' skills, and the improvement in reevaluation scores could simply be due to higher effort. In this section, I examine the impact of assigning teachers to remedial training on the standardized test scores of their students.

Table 4 summarizes the results of the fuzzy RD regressions with student test scores as the dependent variable, and Figure 6 plots the raw data alongside the results of a quadratic (sharp) RD regression. Importantly for internal validity, there is no statistically significant difference in the test scores of students in the year of their teacher's first evaluation, as shown in column 1 of Table 4. During the first three years of remediation (i.e., the three years after teachers are first evaluated), there is no impact of being eligible for remedial training on students' standardized test scores, as shown in columns 2 to 4. In particular, these coefficients are in the range of -0.048 to 0.007 standard deviations, and are robust to controlling for year fixed effects, grade-subject fixed effects, student characteristics (gender and socioeconomic status), and teacher baseline characteristics, as shown in Panel B of Table 4. The impact is also small and statistically insignificant if I pool these three years together, as shown in the second column of Appendix Table A.4.

On the year of the second teaching evaluation, which is also the fourth and final year of remedial training, the students of teachers barely assigned to remediation obtain lower test scores, with a negative estimated effect of around 0.12 standard deviations (column 5). This coefficient is statistically significant at the 5% level, and is barely affected by the inclusion of a rich set of controls, as can be seen when comparing the results of Panel A to those of Panel B. These results should be taken with caution, however, since this coefficient is not statistically significant after accounting for multiple hypothesis testing. To abstract from changes in the grades and subjects

that participate in SIMCE, I also run these regressions for the subsample of $4^{th}$ graders, the only grade level tested on a yearly basis. Appendix Table A.5 shows that the drop in standardized test scores in the year in which teachers are reevaluated is also observed in this subsample, particularly in language. These results are much noisier than the full sample, but this is expected given the considerably smaller sample size.

How can four years of remedial training result in higher teaching evaluation scores, but lower student achievement? If teachers assigned to remedial training face pressure to obtain higher evaluation scores, they may react by putting more effort into preparing for their evaluation. Time constraints might result in these teachers allocating time away from their other duties, as predicted by models of multitasking (Holmstrom and Milgrom, 1991; Baker, 1992, 2002), leading to a drop in student learning.[30] This is not an unlikely hypothesis, since a frequent concern voiced by Chilean teachers regarding their evaluation is the amount of workload required to complete the portfolio (Taut et al., 2011; Taut and Sun, 2014). This concern is exacerbated by the fact that teachers do not receive extra time to work on their assessment, something that is clearly reflected in the questionnaire that accompanies teachers' evaluation, where 80% of respondents reported lack of time as one of the main difficulties in developing their portfolio. Other reasons such as lack of information, problems with being filmed or lack of familiarity with computers were invoked by less than a third of teachers.[31] The median teacher in the sample works 38 contractual hours per week, of which 75% are spent teaching, and the remaining 25% are devoted to non-classroom activities. However, teachers also work during their free time. In a survey to 12,000 teachers conducted in 2012, 60% of respondents reported devoting 10 or more hours a week to work-related tasks outside their working hours (Centro UC, 2016).

Are there any gains from remedial training other than higher scores in teaching evaluations?

[30] This type of behavior has been reported in several studies examining the effect of tying teacher pay to the performance of students, such as Jacob and Levitt, 2003, Figlio and Winicki, 2005, Figlio, 2006, Glewwe et al., 2010, and Behrman et al., 2015.

[31] A team from the OECD that visited Chile to analyze its teaching evaluation system also reported that teachers struggle to find time to adequately prepare and respond to all the requirements of the evaluation (Santiago et al., 2013). The time constraints faced by teachers during evaluation years can also be seen in the fact that, conditioning on school and year fixed effects, $4^{th}$ grade students obtain test scores 1% of a standard deviation lower (statistically significant at the 1%) when their teacher is being evaluated.

One year after remediation is finalized (i.e., five years after the first teacher assessment), the students of teachers assigned to remedial training score 0.11 standard deviations higher than those who have a teacher that did not attend remediation. A similar pattern of results arises when using a higher-order polynomial of the running variable, as shown in Section 6.2. However, these findings are not robust, as the coefficient is only statistically significant at the 10% level, and is not significant at conventional levels after controlling for teacher and student characteristics. Furthermore, the raw data in Figure 6 does not show a clear jump in test scores at the cutoff. Two years after remediation is over, the test scores of students whose teachers were barely assigned to remediation are not statistically different from that of students whose teachers barely got out of it, as can be seen in the last column of Table 4. The impact on student learning is not statistically different from zero if I pool the two years after teachers' reevaluation, as can be seen in the last column of Appendix Table A.4.

# 6 Validity and Robustness Checks

## 6.1 Results are Not Driven by Differential Attrition or Sorting

As shown in detail in Section 3.1, teachers evaluated with a Basic and Competent rating in their first assessment do not systematically differ in terms of their baseline characteristics. However, as my empirical analysis relies on outcomes measured several years after teachers are first evaluated, estimating the causal impact of remediation also requires that these two groups of teachers do not experience differential attrition or sorting. Given that 35% of the teachers first evaluated between 2004 and 2010 had not been reassessed by 2015, a plausible concern is that attrition rates are discontinuous at the threshold, biasing the estimates of the effect of remediation. Mitigating this concern, Panel A of Table 5 shows that the probability of being reevaluated is continuous across the cutoff, as is the number of years between teachers' first and second evaluation.[32] A second potential

[32] Almost half of the attrition is explained by teachers who retired or were close to retirement and thus did not have to take the test (16% of Basic/Competent teachers). The null impact of remediation on other sources of attrition is probably due to the fact that it is quite uncommon for public school teachers to get a job in a private school or quit

threat is that the working conditions of teachers who were barely assigned and not assigned to remedial training may have diverged after their first evaluation, mechanically leading to differences in their reevaluation scores or in the achievement of their students. For example, teachers assigned to remediation could have cut down on their working hours or may have concentrated all of their teaching in just one school to have more time to prepare for their reevaluation. It turns out, however, that at the moment of reevaluation there are no differences in the working conditions of teachers who obtained a Basic and Competent score in their first evaluation, as displayed in Panel B of Table 5. In Appendix B, I provide details on why the attrition rate and job characteristics of teachers who remain in the public school system do not vary at the cutoff between a Basic and a Competent rating.

A point to consider when examining the impact on standardized test scores is whether teachers barely eligible and ineligible for remedial training differ in the likelihood of teaching students who participate in SIMCE, since these tests only cover certain grades and subjects. In fact, only two thirds reevaluated teachers instruct one of these grades and subjects at some point in the six years after their first evaluation. One threat to identifying the causal impact of remedial training on student achievement is that school principals may be less prone to assign teachers who are attending remediation to the grades that participate in SIMCE. It turns out, however, that there is no discontinuity in the likelihood of teaching a SIMCE subject and grade across the Basic/Competent cutoff, as shown in Panel C of Table 5. Teachers at both sides of the threshold are equally likely to instruct one of these subjects/grades in each of the six years after their first evaluation, and the total number of years is continuous.

Within the sample of teachers whose students participate in SIMCE, an additional threat is the possibility that school principals assign teachers who attend remediation to classes with worse or better students, biasing the estimates downwards or upwards, respectively. Since students are not tested on a yearly basis, I cannot check whether students' lagged standardized test scores are

teaching altogether. Only 4% of teachers with a Basic or Competent rating were working in a private school four years after their first assessment, and 9% left the school system or took an administrative position. The remaining 6% were still teaching in a public school but did not participate in their second evaluation by 2015.

smooth across the Basic/Competent cutoff. However, I have information on their yearly GPAs, pass rates, and attendance.[33] If principals were strategically assigning teachers on the basis of students' past performance, these are probably the measures they would use. As shown in Appendix Table A.6, there are hardly any differences in the lagged performance of students for teachers with a Basic and Competent rating in each of the six years after teachers' first evaluation.[34]

The evidence presented above suggests that on a given year, teachers who were barely assigned to remedial training have the same likelihood of teaching a SIMCE grade and subject. And for those who teach one of these grades and subjects, predetermined student characteristics do not differ, lending a causal interpretation to the estimates in Table 4. However, a potential problem when comparing the test score impacts for the different years after teachers are first evaluated is that the sample of teachers is not constant across years. In fact, almost none of these teachers instructs a subject and grade that participates in SIMCE the entire six-year period after their first evaluation, with most teachers only doing so between one and three years. As explained in Section 4, the sample changes across years in Table 4 occur because the assessment for $8^{\text{th}}$ and $10^{\text{th}}$ graders is only carried out every other year for most of the analysis period, and because teachers in Chile usually rotate across grades.[35] Even if teachers and students at both sides of the cutoff are comparable on a given year, a possible drawback of comparing results across years is that teachers who instruct a SIMCE grade and subject the first year after their evaluation may be different from those who do so four years after, for example, and these differences may interact with how assignment to remedial training affects student performance. I investigate the relevance of this concern by comparing the baseline characteristics of teachers in these six subsamples in Panel A of Appendix Table A.7. Although most of these characteristics are relatively constant, there is a striking differ-

[33] Although these measures are imperfect proxies of student ability, they are strongly correlated with standardized test scores. In $4^{\text{th}}$, $8^{\text{th}}$ and $10^{\text{th}}$ grade, students with a GPA that is one standard deviation higher obtain, on average, SIMCE scores 79% of a standard deviation higher in that same year. Furthermore, those who passed the school year obtained SIMCE scores 74% of a standard deviation higher during that year, and students with a standard deviation higher attendance rate obtain SIMCE scores 15% of a standard deviation higher.

[34] The only statistically significant difference is in the fifth year after teachers' first evaluation, where students of teachers assigned to remediation have a significantly higher attendance rate. However, this difference is tiny (1 percentage point over a mean attendance of 91%).

[35] The sample also changes across years because the optimal bandwidth varies. Importantly, as shown in Section 6.2, the results in Table 4 are robust to reasonable bandwidth modifications.

ence across subsamples in the total number of teachers, and the grade level at which they taught when they were first evaluated (i.e., lower primary, upper primary or secondary school).[36] Because teacher evaluations were gradually rolled out across municipalities, subjects and education levels, there is considerable variation in the number of evaluated teachers across years.[37] Together with the fact that $8^{th}$ and $10^{th}$ graders participate in SIMCE every other year, this generates differences in the number of teachers and their levels of instruction across the different subsamples.[38] However, if I focus only on teachers who instruct $4^{th}$ grade, the grade that participates in SIMCE every year, teachers are very similar in terms of their observable characteristics across subsamples (Panel B of Appendix Table A.7), and the number of teachers is constant across years. Given that teachers appear to be comparable across years, it is reasonable to assume, for example, that the drop and increase in student learning on the year of reevaluation and the following year are not driven by the characteristics of teachers in each of these samples. Further mitigating this concern, Appendix Figure A.11 shows that the impact of assignment to remediation on teachers' reevaluation scores does not differ in a systematic way across these different subsamples.[39]

## 6.2 Robustness Checks

Choosing a bandwidth in an RD estimation implies a tradeoff between bias and precision. While larger bandwidths lead to more precise estimates, they also increase the risk of bias. Figures

[36] Lower primary school ranges from $1^{st}$to $4^{th}$ grade, upper primary school from $5^{th}$ to $8^{th}$, and secondary school from $8^{th}$ to $12^{th}$.

[37] In 2004, for example, only 107 municipalities participated in the evaluation, and only math, language and science teachers in primary school were assessed (1,719 teachers). In 2005, the evaluation incorporated the remaining municipalities, and also covered secondary school math and language teachers, reaching 10,631 teachers in total. Since most teachers had already been evaluated at least once by 2008, the number of evaluated teachers dropped to 7,685 and 3,609 in 2009 and 2010, respectively.

[38] Take for example teachers who instruct math in upper primary school ($5^{th}$ to $8^{th}$ grade). Around 600 of these teachers were first evaluated in 2005, but almost 3,000 were assessed the year after. Since $8^{th}$ graders participated in SIMCE in 2007, 2009 and 2011, the few $8^{th}$ grade teachers evaluated in 2005 will have their students assessed 2, 4 and 6 years after their first evaluation, and the almost 3,000 teachers evaluated in 2006 will have their students assessed 1, 3 and 5 years after their first assessment. Thus, the total number of $8^{th}$ grade math teachers in these different samples will vary a lot.

[39] Although the coefficients are more imprecisely estimated than those of the full sample of reevaluated teachers, the point estimates have the same sign and a similar magnitude in most cases. The full sample is composed of 37,866 reevaluated teachers, of which 13,161 never teach a SIMCE subject/grade in the six-year period. The number of teachers in each of the other subsamples is much smaller, ranging from 8,138 to 10,740 teachers.

A.12 and A.13 plot the fuzzy RD point estimates of the impact of remedial training on teachers' reevaluation scores and their students' test scores, respectively, along with their 95% conventional confidence interval for a wide range of bandwidths. The solid line indicates the MSE-optimal bandwidth, which I use in my baseline specification. Importantly, my estimates are robust to using different bandwidths, although they become quite imprecise for arbitrarily small bandwidths.

Another important decision in an RD design is the choice of the local polynomial order. For a given bandwidth, using a polynomial of higher order improves the accuracy of the approximation but also the variability of the treatment effect estimator (Cattaneo et al., 2019). Although my baseline specification is a local linear regression, I show in Appendix Tables A.8 and A.9 that my results are robust to using a second order polynomial instead.

# 7 Discussion

## 7.1 Is the Improvement in Reevaluation Scores Driven by Effort or Training?

Singling low-performing teachers out for remedial training may impose a stigma which could possibly lead them to exert more effort. Some of the findings in this paper support this hypothesis. In particular, the most plausible explanation for the negative impact on student test scores on the year of teachers' reevaluation is that low-performing teachers face the stigma of being assigned to remediation, put more effort into preparing for their evaluation, and thus dedicate less time to their other duties. To examine whether teachers assigned to remediation indeed experienced stigma, I use responses from a written survey to all evaluated teachers. Teachers hand in their responses together with their portfolio, and are informed that these surveys are confidential and have no bearing on the results of their assessment. Reevaluated teachers are asked how they were affected by the results of their previous evaluation, and what actions that they took as a consequence. As shown in columns 5-6 of Table 6, teachers barely assigned to remedial training were 31 percentage points less likely to report that their prestige and job satisfaction increased as a consequence of

their first evaluation results.[40] These estimates are significant at the 1% level, and very robust to the inclusion of year fixed effects and baseline controls. These results are in line with the findings of Koedel et al. (2017), who also show that teachers with higher ratings in their teaching evaluations in Tennessee were subsequently more satisfied with their job. Teachers with a Basic score were also less likely to say that their job stability went up (10 percentage points), that the responsibilities assigned to them increased (7 percentage points), or that their income went up (31 percentage points), providing strong evidence of the existence of a stigma associated with being assigned to remediation. As shown in columns 2 and 4, teachers who received a Basic rating were also more likely to report that they asked for support to interpret the results of their first evaluation (9 percentage points), and met with the principal to discuss these results (4 percentage points). These results suggest that teachers who were barely assigned to remedial training put more effort into improving their assessment results.

However, it is unclear whether the improvement in teaching evaluations is only explained by higher effort, or if the skills teachers acquire in remedial training also play a role. Supporting the latter hypothesis, teachers assigned to remediation only had higher reevaluation scores in the areas which were covered during remediation, such as lesson planning, student assessment and reflection about teaching practices. Furthermore, the average teacher who attended remediation gave the training a rating of 5 out of 7 in terms of quality, relevance and usefulness, and 67% expected their participation in these activities to lead to higher reevaluation scores, as shown in Appendix Table A.1. A way of assessing whether remediation had a direct impact is to test for heterogeneous effects by the average attendance to remedial training. As explained in Section 2, despite teachers' obligation to participate in remediation, some teachers did not take part in these activities. If the improvement in reevaluation scores was completely driven by stigma and higher effort, the impact of being assigned to remediation should not vary by the share of teachers who effectively participated in these activities. For each reevaluation year, I split the sample by whether

[40]Only teachers who were reevaluated in 2010-2014 (79% of reevaluated teachers) were asked how they were affected by the results of their first evaluation. Importantly, response rates were high (91%), and do not differ for individuals who barely got a Basic and Competent score in their first evaluation, as can be seen in the last row of Table 6.

the average attendance to remedial training in the teachers' municipality was below or above the median. Average participation in remediation was 63% in the former, and 95% in the latter.[41] As can be seen in the Table 7, the improvement in portfolio scores is concentrated in municipalities and years in which participation in remedial training was high. Taken together, the evidence presented in this section suggests that the improvement in reevaluation scores was driven both by higher effort and skill acquisition during training.

It is important to understand why remedial training led to improvements in teachers' reevaluation scores but did not have lasting impacts on student learning. The findings from a recent meta-analysis of 33 teacher training programs in low- and middle-income countries by Popova et al. (2018) shows that the impacts on student learning are higher when these programs include face-to-face training, incorporate lesson enactment where teachers can practice with one another, focus on subject-specific pedagogy, and link participation to career incentives. A meta-analysis of teacher training programs in high-income countries by Fryer (2017) finds that the most beneficial programs had precise training and curriculum materials for implementers to follow. These results are confirmed by Leme et al. (2012), who find that a comprehensive intervention providing teacher training, curricular organization and pedagogical materials in Brazil led to higher student learning. Although the majority of remedial training was face-to-face (only 1% were online courses), as shown in Appendix Table A.1, and the outcome of remediation was somewhat linked to career incentives, it was very rare for remediation to take the form of role-playing (1%). It is unclear whether remediation activities focused on pedagogical skills related to the subject teachers instruct or had a more general approach. Furthermore, I do not have information on whether teachers were provided with pedagogical materials to complement the remediation activities. It is thus possible that the design of these activities was not suitable to generate profound transformations in teachers' skills. Alternatively, the courses may have just been geared towards improving teachers' evaluation scores (i.e., "training for the test").

[41] I had to exclude teachers who were reevaluated in 2008-2009 from this analysis, as the survey did not include the question on attendance to remediation in those years.

## 7.2 What Role do Differences in Salaries Play?

As discussed in Section 3, the parameter of interest measures the joint impact of being assigned to remedial training and being ineligible for AVDI, a temporary salary increase. The effect of remedial training is thus confounded with the fact that, as compared to teachers who got a Basic rating and were assigned to remediation, some of the teachers who started out with a Competent rating obtained the AVDI bonus. In particular, 47% of the teachers who barely got a Competent rating received the AVDI award. This implied that on average, the salary of teachers who barely got a Basic rating was 1.5% lower than that of teachers who barely got a Competent rating. Although these salary differences are small, it is important to discard that they are driving the results.

To examine whether the results are driven by differential salary increases, one could take advantage of the AVDI assignment process. Teachers who obtain a Competent score and apply for AVDI must take a content mastery test, and only those who obtain a score above a certain threshold obtain this raise, which lasts until their next teaching evaluation. Unfortunately, there is evidence of significant sorting to the left of this threshold, and so it is not possible to provide causal estimates of the impact of receiving the AVDI raise.[42] As an alternative, I assume that for teachers with a Basic rating, the salience of not obtaining this raise increases with the share of peers who received the AVDI. I take advantage of the data contained in the teacher survey to test the validity of this assumption. In schools in which the share of teachers receiving the AVDI was above the median, teachers barely assigned to remedial training were 37 percentage points less likely to report that their first evaluation results led to higher income (vs. teachers in the same set of schools who obtained a Competent rating).[43] In schools where the share of teachers receiving the AVDI was below the median, this point estimate was only 23 percentage points, and both point estimates are statistically different from each other. I then estimate whether the effect of remediation on

[42] The point estimate of the difference in log heights at this threshold is very large and statistically different from zero (-0.890 with a s.e. of 0.047).

[43] During the year of reevaluation (i.e., the year of this survey), 23% of these teachers' peers were receiving the AVDI award on average. For teachers in schools above and below the median, the average was 30% and 13%, respectively. I split the sample within evaluation years and municipalities to keep the quality of remedial training and the difficulty of the evaluation constant across both groups.

teachers' reevaluation scores varies with the share of peers who receive the AVDI. Although the salience of not obtaining the AVDI raise differs for teachers in both types of schools, the impact of obtaining a Basic rating on reevaluation scores is remarkably similar, as shown in Table 8. While this test is not definite, it suggests that the higher reevaluation scores of teachers who were barely assigned to remedial training are at best partially explained by salary differences.

# 8 Conclusion

This paper investigates the impact of remedial training for low-performing teachers in the context of Chile, where teachers with low scores in their teaching evaluations are assigned to four years of remediation. Estimation results relying on a fuzzy RD design indicate that remedial training improves teachers' pedagogical skills as measured by their reevaluation scores, and by the standardized test scores of their students the year after completing remedial training. However, these last results are not robust, and disappear the following year. Perhaps more importantly, during the year of their teachers' reevaluation, the students whose teachers were barely assigned to remedial training experience a drop in their standardized test scores. Taken together, these results indicate that the Chilean remedial training program is not successful at generating meaningful impacts on student learning. It should be noted, however, that the conclusions of this paper only apply to teachers with evaluation scores close to the margin between a Basic and a Competent rating. It is unclear if teachers with a lower performance stand to benefit more or less from remedial training.

# References

**Alvarado, Macarena, Gustavo Cabezas, Denise Falck, and María Elena Ortega**, "La Evaluación Docente y Sus Instrumentos: Discriminación del Desempeño Docente y Asociación con los Resultados de los Estudiantes," Technical Report, Centro de Estudios Ministerio de Educación de Chile y Programa de Desarrollo de las Naciones Unidas [PNUD], Santiago 2012.

**Araujo, M Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady**, "Teacher Quality and Learning Outcomes in Kindergarten," *Quarterly Journal of Economics*, 2016, *131* (3), 1415–1453.

**Baker, George**, "Distortion and Risk in Optimal Incentive Contracts," *Journal of Human Resources*, 2002, pp. 728–751.

**Baker, George P**, "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 1992, pp. 598–614.

**Behrman, Jere R, Susan W Parker, Petra E Todd, and Kenneth I Wolpin**, "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools," *Journal of Political Economy*, 2015, *123* (2), 325–364.

**Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane**, "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa," *Journal of Economic Perspectives*, 2017, *31* (4), 185–204.

**Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff**, "Explaining the Short Careers of High-Achieving Teachers in Schools with Low-Performing Students," *American Economic Review*, May 2005, *95* (2), 166–171.

**Bravo-Urrutia, David, Denise Falck, and Claudia Peirano-Rodríguez**, "Encuesta Longitudinal Docente 2005: Análisis y Principales Resultados," Technical Report, Universidad de Chile, Departamento de Economía 2008.

**Bruns, Barbara and Javier Luque**, *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*, World Bank Publications, 2015.

**_ , Soledad De Gregorio, and Sandy Taut**, "Measures of Effective Teaching in Developing Countries," 2016. Research on Improving Systems of Education (RISE) Working Paper 16/009.

**Cabezas, Verónica, Ricardo Paredes, Francisca Bogolasky, Rosario Rivero, and Magdalena Zarhi**, "First Job and the Unequal Distribution of Primary School Teachers: Evidence for the Case of Chile," *Teaching and Teacher Education*, 2017, *64*, 66–78.

**Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik**, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 2014, *82* (6), 2295–2326.

**Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik**, *A Practical Introduction to Regression Discontinuity Designs: Foundations* Elements in Quantitative and Computational Methods for the Social Sciences, Cambridge University Press, 2019.

**Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas**, *Marco para la Buena Enseñanza*, 7 ed. 2008.

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

**Centro de Políticas Públicas UC and Elige Educar**, "Uso del Tiempo No Lectivo: Desafíos para Políticas Públicas y Comunidades Educativas," Technical Report 2016.

**Figlio, David N**, "Testing, Crime and Punishment," *Journal of Public Economics*, 2006, *90* (4), 837–851.

_ **and Joshua Winicki**, "Food for Thought: The Effects of School Accountability Plans on School Nutrition," *Journal of Public Economics*, 2005, *89* (2), 381–394.

**Fryer, Roland G**, "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments," in "Handbook of Economic Field Experiments," Vol. 2, Elsevier, 2017, pp. 95–322.

**Glewwe, Paul, Nauman Ilias, and Michael Kremer**, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–227.

**Hanushek, Eric A and Steven G Rivkin**, "Generalizations about Using Value-Added Measures of Teacher Quality," *American Economic Review*, 2010, *100* (2), 267–271.

_ **, John F Kain, and Steven G Rivkin**, "Why Public Schools Lose Teachers," *Journal of Human Resources*, 2004, *39* (2), 326–354.

**Holmstrom, Bengt and Paul Milgrom**, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, 1991, *7*, 24–52.

**Hoyos, Rafael E De, Alejandro Jorge Ganimian, and Peter A Holland**, "Teaching with the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina," 2017. World Bank Policy Research Working Paper 8261.

**Jacob, Brian A and Lars Lefgren**, "The Impact of Teacher Training on Student Achievement. Quasi-Experimental Evidence From School Reform Efforts in Chicago," *Journal of Human Resources*, 2004, *39* (1), 50–79.

_ **and Steven D Levitt**, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 2003, pp. 843–877.

**Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger**, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," Technical Report, Bill and Melinda Gates Foundation 2013.

**_ , Eric S Taylor, John H Tyler, and Amy L Wooten**, "Identifying Effective Classroom Practices Using Student Achievement Data," *Journal of Human Resources*, 2011, *46* (3), 587–613.

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

**Kraft, Matthew A**, "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies," *Journal of Human Resources*, 2019, *54* (1), 1–36.

**Lee, David S and Thomas Lemieux**, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

**Leme, Maria Carolina, Paula Louzano, Vladimir Ponczek, and André Portela Souza**, "The Impact of Structured Teaching Methods on the Quality of Education in Brazil," *Economics of Education Review*, 2012, *31* (5), 850–860.

**Manzi, Jorge, Roberto González, Yulan Sun, Rodolfo Bonifaz, María Paulina Flotts, Andrea Abarzúa, Paulina Calderón, Nelson Valerion, Pablo Torres, Mónica Correa et al.**, *La Evaluación Docente en Chile*, MIDE UC, 2011.

**McCrary, Justin**, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

**Mizala, Alejandra and Pilar Romaguera**, "School Performance and Choice: the Chilean Experience," *Journal of Human Resources*, 2000, pp. 392–417.

**Muralidharan, Karthik and Venkatesh Sundararaman**, "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India," *Economic Journal*, 2010, *120* (546), F187–F203.

**OECD**, *TALIS 2013 Results: An International Perspective on Teaching and Learning*, OECD Publishing, 2014.

**Ortúzar, María Soledad, Pamela Ayala, Carolina Flores, and Carolina Milesi**, "Percepciones Acerca del Proceso de Búsqueda y Contratación de Docentes en Chile: Nudos Críticos e Inequidad del Sistema," *Calidad en la educación*, 2016, (45), 251–287.

**Paredes, R, F Bogolasky, V Cabezas, R Rivero, and M Zahri**, "Los Determinantes del Primer Trabajo para Profesores de Educación Básica en la Región Metropolitana," Technical Report 2013.

**Popova, Anna, David K. Evans, Mary E. Breeding, and Violeta Arancibia**, "Teacher Professional Development around the World : The Gap between Evidence and Practice," 2018. World Bank Policy Research Working Paper 8572.

**Rockoff, Jonah E**, "The Impact of Individual Teachers on Student Achievement: Evidence From Panel Data," *American Economic Review*, 2004, *94* (2), 247–252.

**Santiago, Paulo, Francisco Benavides, Charlotte Danielson, Laura Goe, Deborah Nusche et al.**, "Teacher Evaluation in Chile," *OECD Reviews of Evaluation and Assessment in Education. Santiago de Chile: OCDE*, 2013.

**Scafidi, Benjamin, David L Sjoquist, and Todd R Stinebrickner**, "Race, Poverty, and Teacher Mobility," *Economics of Education Review*, 2007, *26* (2), 145–159.

**Taut, Sandy and Yulan Sun**, "The Development and Implementation of a National, Standards-Based, Multi-Method Teacher Performance Assessment System in Chile," *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, 2014, *22*.

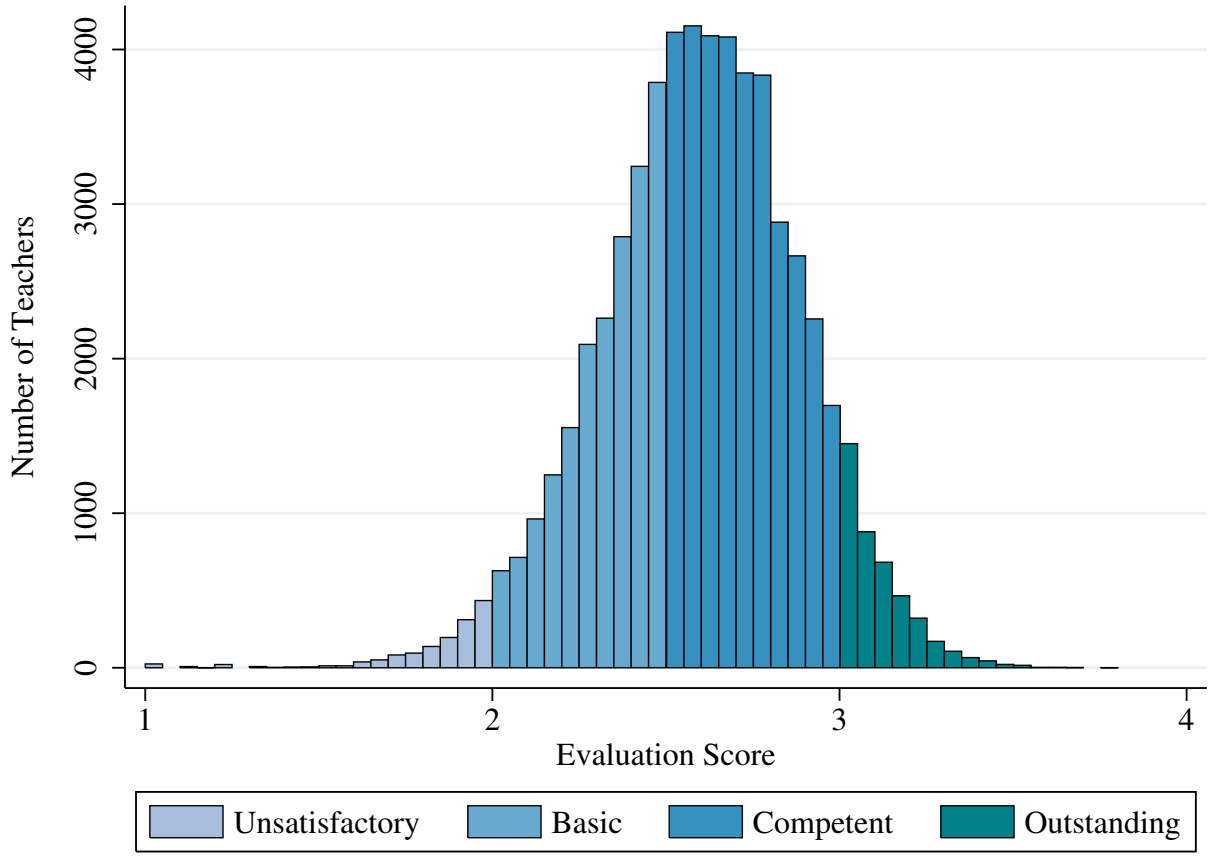
_ , **Edgar Valencia, Diego Palacios, Maria V. Santelices, Daniela Jim´nez, and Jorge Manzi**, "Teacher Performance and Student Learning: Linking Evidence from Two National Assessment Programmes," *Assessment in Education: Principles, Policy & Practice*, 2016, *23* (1), 53–74.

_ , **Maria Verónica Santelices, Carolina Araya, and Jorge Manzi**, "Perceived Effects and Uses of the National Teacher Evaluation System in Chilean Elementary Schools," *Studies in Educational Evaluation*, 2011, *37* (4), 218–229.

**Taylor, Eric S and John H Tyler**, "The Effect of Evaluation on Teacher Performance," *American Economic Review*, 2012, *102* (7), 3628–51.

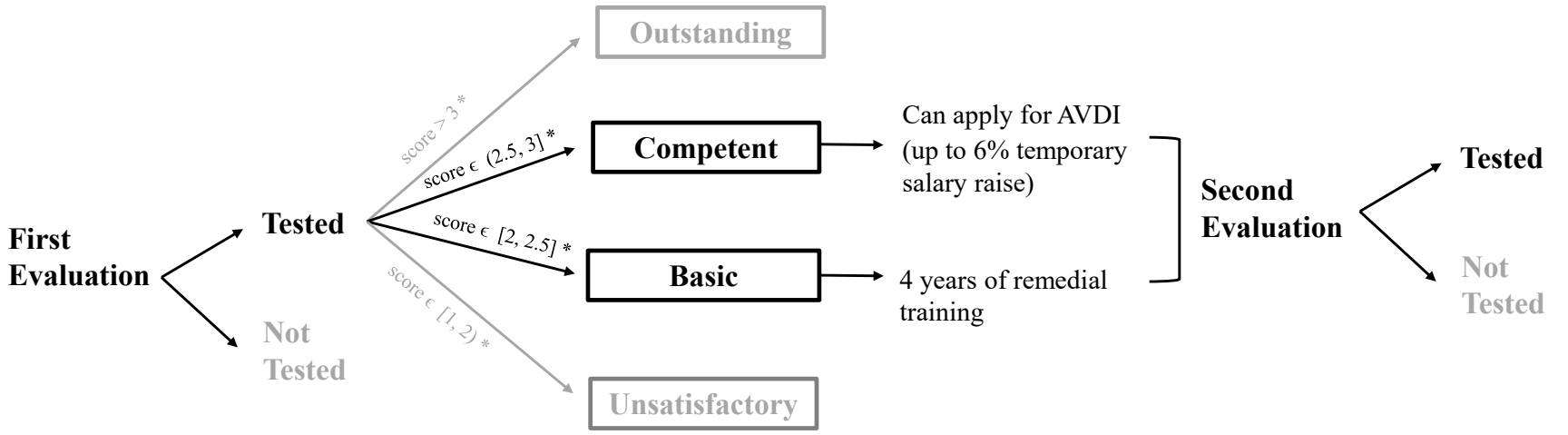
**Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L Shapley**, "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033.," *Regional Educational Laboratory Southwest (NJ1)*, 2007.

Figure 1: Distribution of Evaluation Scores for Teachers First Evaluated in 2004-2010



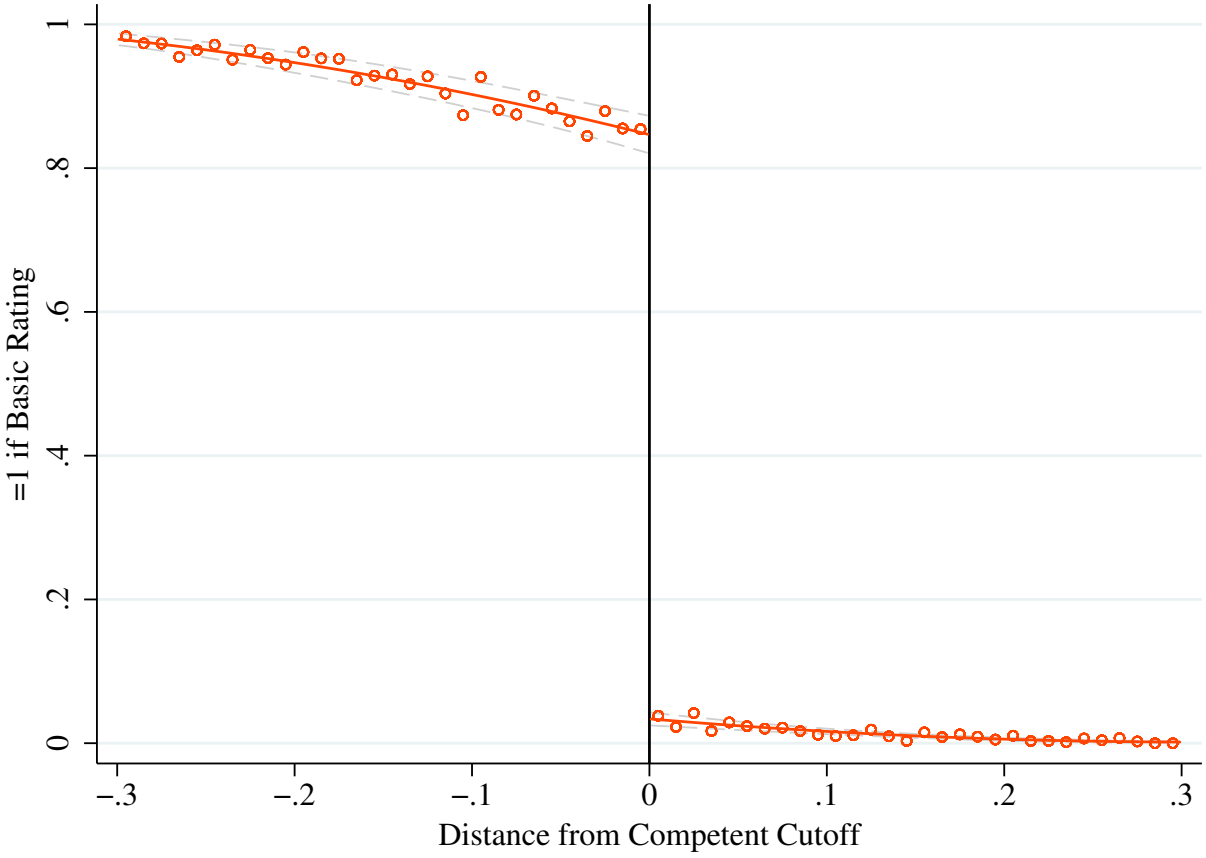

*Notes*: The sample includes all public school teachers instructing primary and secondary school who were evaluated for the first time in 2004-2010. Scores are broken down into four categories based on fixed cutoff rules. Final categories may differ due to modifications by municipal evaluation committees.

Figure 2: Evaluation Process and Consequences for Teachers First Evaluated in 2004-2010




** Final categories can be modified by the municipal evaluation committee that convenes at the start of $t_0$+1*

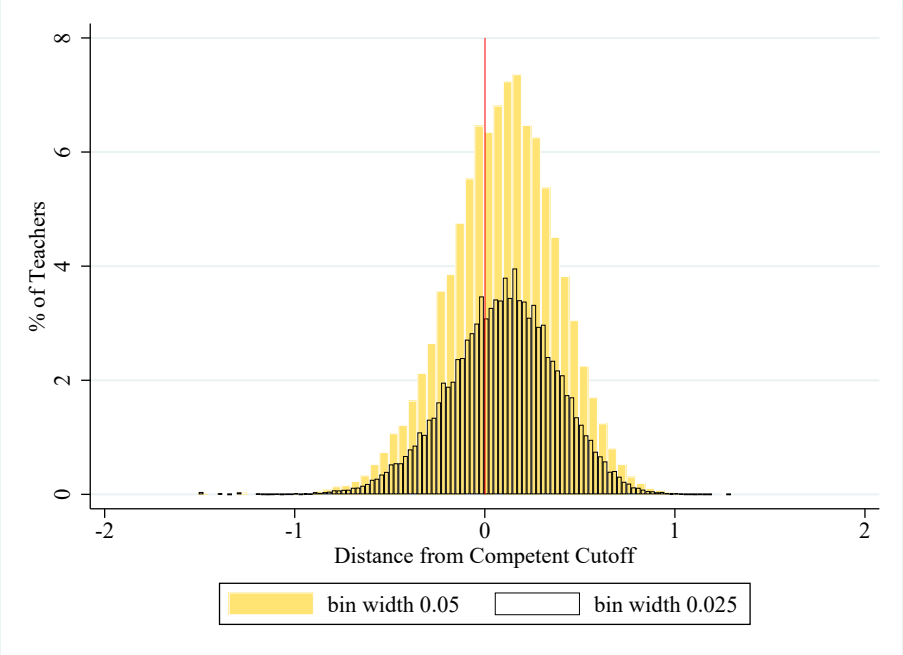
Figure 3: Jump in Probability of Obtaining a Basic Score for Teachers First Evaluated in 2004-2010



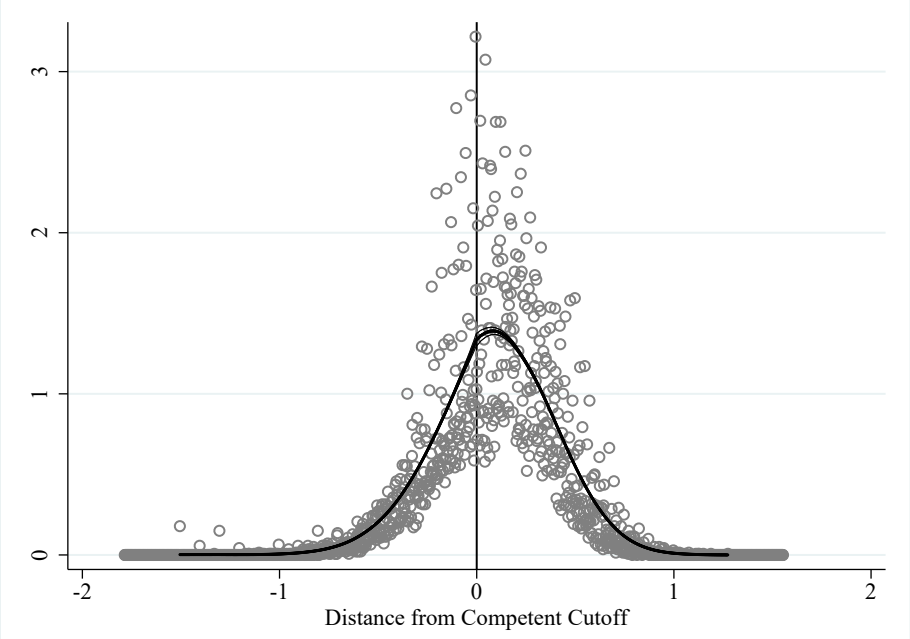

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. The running variable is the evaluation score centered around the Basic/Competent threshold, and the outcome variable is a dummy variable taking the value of 1 if the individual's final rating (after revision by the municipal evaluation committee) is Basic. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality.

Figure 4: Evaluation Scores for Teachers First Evaluated in 2004-2010

Panel A: Histogram of Evaluation Scores

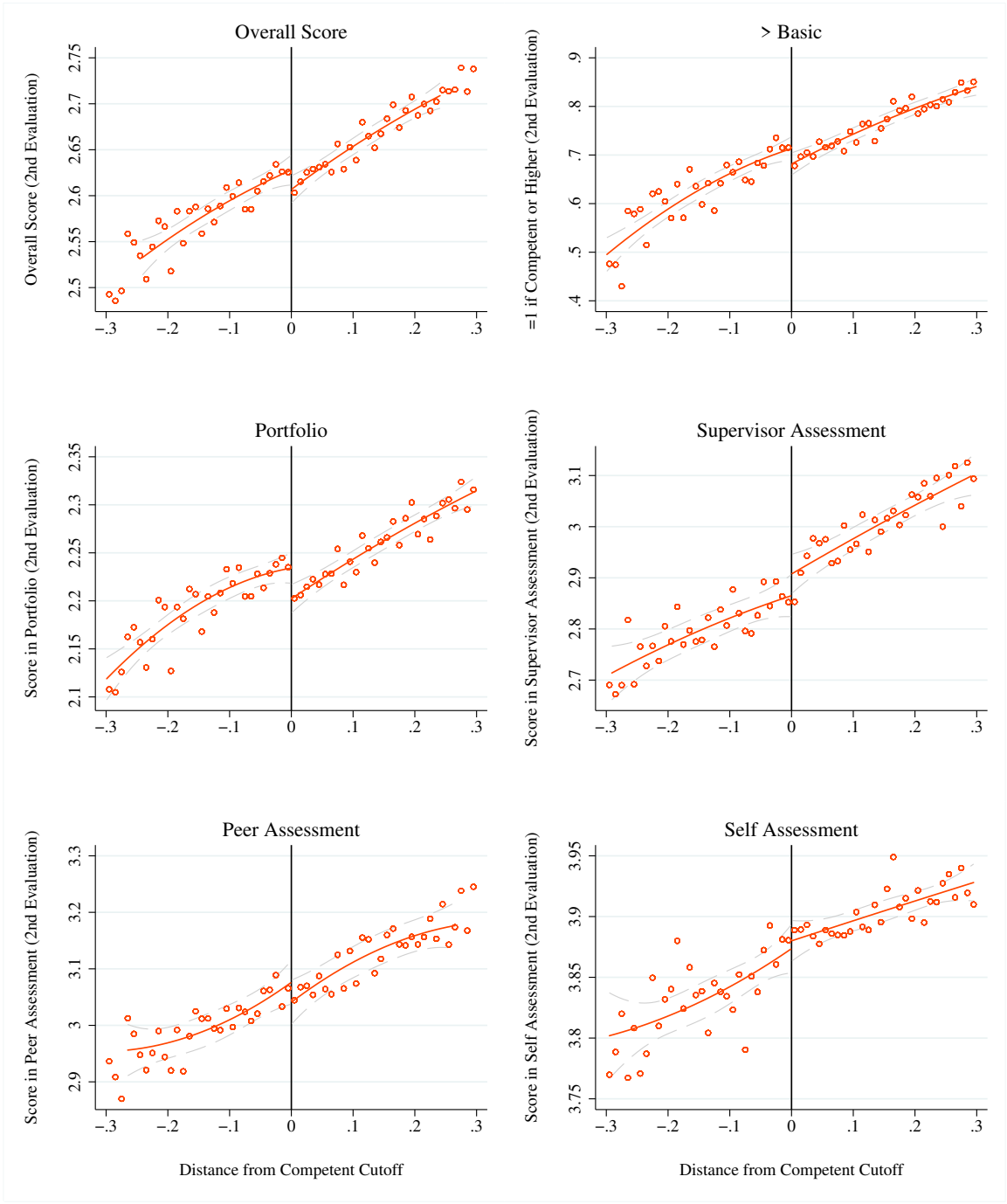



Panel B: Density of Evaluation Scores (McCrary Test)



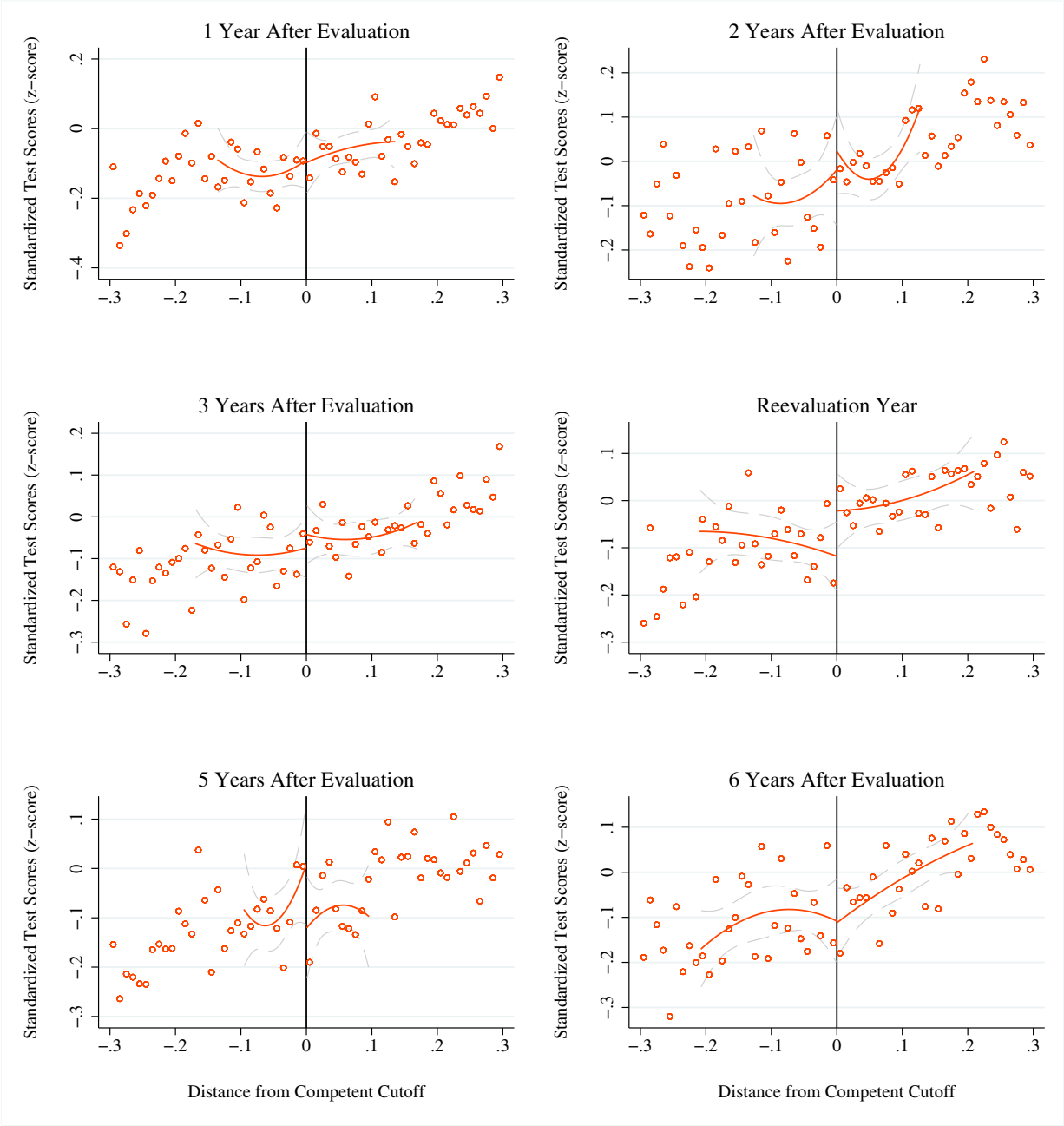

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. The top figure plots the histogram of the running variable (i.e., the evaluation score centered around the Basic/Competent threshold), with bins of 0.05 and 0.025 points. The bottom graph plots the weighted kernel estimation of the log density of the running variable, performed separately on either side of the Basic/Competent cutoff, created using the "DCdensity" Stata command written by Justin McCrary. The point estimate of the difference in the log heights at the threshold is -0.006 (s.e. 0.021).

Figure 5: Effect of Assignment to Remediation on Reevaluation Results




*Notes*: The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The running variable is the first evaluation score centered around the Basic/Competent threshold, and the outcome variable is the result of the second evaluation in one of six dimensions. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and *> Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer Assessment*, *Supervisor Assessment* and *Self Assessment* are the scores in the peer, supervisor and self assessment, respectively.

Figure 6: Effect of Assignment to Remediation on Students' Standardized Test Scores




*Notes*: The sample is composed of all 4th, 8th and 10th grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, and only consider cases in which the students were tested between one and six years after the teacher's first evaluation. The running variable is the teachers' first evaluation score centered around the Basic/Competent threshold, and the outcome variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year). Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the student's school.

Table 1: Balance in Covariates for Teachers Evaluated at the Basic vs. Competent Level in 2004-2010

| | Coefficient | Standard Error | Mean |
|---|---|---|---|
| Teacher and Job Characteristics | | | |
| Male | -0.026 | (0.016) | 0.310 |
| Age | 0.655 | (0.492) | 45.968 |
| Years of experience | 0.791 | (0.562) | 18.277 |
| Has a degree | -0.001 | (0.008) | 0.961 |
| Main job is teaching | -0.006 | (0.008) | 0.965 |
| More than one school | -0.004 | (0.016) | 0.147 |
| More than one municipality | 0.000 | (0.013) | 0.100 |
| Civil servant | -0.011 | (0.020) | 0.688 |
| Number of hours | 0.314 | (0.399) | 34.332 |
| Refused evaluation before | 0.001 | (0.007) | 0.026 |
| Evaluated Subject and Level | | | |
| Math | 0.002 | (0.013) | 0.127 |
| Language | -0.009 | (0.011) | 0.132 |
| Social science | -0.004 | (0.010) | 0.093 |
| Natural science | -0.006 | (0.009) | 0.091 |
| English | -0.007 | (0.007) | 0.045 |
| Art or music | -0.001 | (0.006) | 0.036 |
| Physical Education | -0.009 | (0.007) | 0.058 |
| Other Subjects | 0.007 | (0.009) | 0.062 |
| Lower primary school | 0.015 | (0.020) | 0.356 |
| Upper primary school | -0.007 | (0.020) | 0.412 |
| Secondary school | -0.011 | (0.020) | 0.231 |
| School Characteristics | | | |
| Urban | 0.025 | (0.034) | 0.755 |
| Ln(Enrollment) | -0.029 | (0.093) | 5.975 |
| Teacher-student ratio | 0.050 | (0.827) | 30.219 |
| Average SES of students | 0.012 | (0.073) | 2.028 |
| School average SIMCE | 0.863 | (1.941) | 237.934 |
| School won previous SNED | 0.047** | (0.023) | 0.293 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. Every row shows the result of a separate fuzzy RD using a local linear regression, where the dependent variable is the covariate specified in the row header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. The first column presents the fuzzy RD bias-corrected coefficient, and the second presents the robust bias-corrected standard errors, clustered by the teacher's municipality. The third column presents the mean value of the dependent variable. All the dependent variables are measured at the year of the teachers' first evaluation, in the school in which they were evaluated, except for the school's average SIMCE score which is lagged by one year. The regressions where the dependent variable is a teacher characteristics drop 4% of teachers for which these data were missing, and those with school characteristics drop 0.3% of teachers without a school identifier. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 2: Summary Statistics for Teachers First Evaluated in 2004-2010

| | Full Sample | | | Reevaluated | | | Reev.+ SIMCE | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | N | Mean | SD | N | Mean | SD | N |
| Teacher and Job Characteristics | | | | | | | | | |
| Male | 0.31 | 0.46 | 56,181 | 0.32 | 0.47 | 36,302 | 0.28 | 0.45 | 17,176 |
| Age | 45.97 | 10.36 | 56,154 | 44.32 | 9.34 | 36,297 | 44.40 | 9.29 | 17,173 |
| Years of experience | 18.28 | 11.90 | 56,181 | 16.46 | 10.84 | 36,302 | 16.78 | 11.05 | 17,176 |
| Has a degree | 0.96 | 0.19 | 56,181 | 0.96 | 0.19 | 36,302 | 0.98 | 0.16 | 17,176 |
| Main job is teaching | 0.96 | 0.18 | 56,181 | 0.97 | 0.18 | 36,302 | 0.95 | 0.21 | 17,176 |
| More than one school | 0.15 | 0.35 | 56,181 | 0.15 | 0.36 | 36,302 | 0.10 | 0.30 | 17,176 |
| More than one municipality | 0.10 | 0.30 | 56,181 | 0.11 | 0.31 | 36,302 | 0.06 | 0.25 | 17,176 |
| Civil servant | 0.69 | 0.46 | 56,181 | 0.68 | 0.47 | 36,302 | 0.70 | 0.46 | 17,176 |
| Number of contract hours | 34.33 | 7.53 | 56,181 | 34.64 | 7.39 | 36,302 | 35.89 | 6.09 | 17,176 |
| School Characteristics | | | | | | | | | |
| Lower primary school | 0.36 | 0.48 | 58,585 | 0.37 | 0.48 | 37,866 | 0.48 | 0.50 | 17,953 |
| Upper primary school | 0.41 | 0.49 | 58,585 | 0.40 | 0.49 | 37,866 | 0.41 | 0.49 | 17,953 |
| Secondary school | 0.23 | 0.42 | 58,585 | 0.23 | 0.42 | 37,866 | 0.11 | 0.31 | 17,953 |
| Urban | 0.75 | 0.43 | 58,392 | 0.73 | 0.44 | 37,761 | 0.65 | 0.48 | 17,905 |
| Enrollment | 608.70 | 517.75 | 58,279 | 590.89 | 504.84 | 37,692 | 496.13 | 435.78 | 17,875 |
| Teacher-Student ratio | 30.22 | 9.17 | 58,273 | 29.82 | 9.36 | 37,690 | 28.29 | 9.75 | 17,875 |
| Average SES of students | 2.03 | 0.78 | 58,308 | 1.99 | 0.78 | 37,705 | 1.95 | 0.77 | 17,876 |
| School average SIMCE | 237.93 | 23.43 | 54,050 | 237.75 | 22.93 | 34,802 | 237.05 | 21.08 | 16,188 |
| School won previous SNED | 0.29 | 0.46 | 58,392 | 0.30 | 0.46 | 37,761 | 0.29 | 0.45 | 17,905 |
| First Evaluation Results and Reev. Rates | | | | | | | | | |
| Refused evaluation before | 0.03 | 0.16 | 58,585 | 0.02 | 0.15 | 37,866 | 0.03 | 0.16 | 17,953 |
| Final score (1-4) | 2.60 | 0.29 | 58,585 | 2.61 | 0.29 | 37,866 | 2.62 | 0.26 | 17,953 |
| Basic category or below | 0.34 | 0.47 | 58,585 | 0.32 | 0.47 | 37,866 | 0.31 | 0.46 | 17,953 |
| Committee modified category | 0.05 | 0.21 | 58,585 | 0.05 | 0.21 | 37,866 | 0.04 | 0.20 | 17,953 |
| Portfolio score (1-4) | 2.23 | 0.33 | 58,585 | 2.25 | 0.33 | 37,866 | 2.24 | 0.30 | 17,953 |
| Peer assessment score (1-4) | 2.88 | 0.66 | 58,572 | 2.89 | 0.66 | 37,857 | 2.92 | 0.65 | 17,946 |
| Self assessment score (1-4) | 3.79 | 0.46 | 58,584 | 3.80 | 0.45 | 37,866 | 3.78 | 0.45 | 17,953 |
| Supervisor assessment score (1-4) | 3.08 | 0.69 | 58,448 | 3.10 | 0.67 | 37,776 | 3.17 | 0.65 | 17,903 |
| Reevaluated | 0.65 | 0.48 | 58,585 | 1.00 | 0.00 | 37,866 | 1.00 | 0.00 | 17,953 |
| Years until reevaluated | 4.29 | 0.89 | 37,866 | 4.29 | 0.89 | 37,866 | 4.00 | 0.00 | 17,953 |
| Taught a SIMCE grade/subject | | | | | | | | | |
| 1-6 years after evaluation | 0.57 | 0.50 | 58,585 | 0.65 | 0.48 | 37,866 | 1.00 | 0.00 | 17,953 |
| 1 year after evaluation | 0.26 | 0.44 | 58,585 | 0.27 | 0.44 | 37,866 | 0.40 | 0.49 | 17,953 |
| 2 years after evaluation | 0.19 | 0.39 | 58,585 | 0.21 | 0.41 | 37,866 | 0.35 | 0.48 | 17,953 |
| 3 years after evaluation | 0.23 | 0.42 | 58,585 | 0.28 | 0.45 | 37,866 | 0.43 | 0.50 | 17,953 |
| 4 years after evaluation | 0.17 | 0.37 | 58,585 | 0.23 | 0.42 | 37,866 | 0.38 | 0.48 | 17,953 |
| 5 years after evaluation | 0.21 | 0.40 | 58,585 | 0.28 | 0.45 | 37,866 | 0.43 | 0.50 | 17,953 |
| 6 years after evaluation | 0.17 | 0.37 | 58,585 | 0.24 | 0.43 | 37,866 | 0.37 | 0.48 | 17,953 |
| Number of years (1-6) | 1.21 | 1.37 | 58,585 | 1.51 | 1.45 | 37,866 | 2.35 | 1.19 | 17,953 |

*Notes:* The sample in the first three columns is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010, whereas the sample in columns (4)-(6) is restricted to teachers who were also reevaluated by 2015. The sample in columns (7)-(9) is further restricted to teachers who were reevaluated after 4 years and also taught a grade/subject that was evaluated in SIMCE one to six years after their first evaluation. All of the teacher and school characteristics are measured at the year of the teachers' first evaluation, in the school in which they were evaluated, except for the school's average SIMCE score which is lagged by one year. *More than one school* and *More than one municipality* are dummies for whether the teacher worked in more than one school or municipality in the year of the first evaluation. *Civil servant* is a dummy variable taking the value of 1 if the teacher has a civil servant position, and 0 if he/she is a contract teacher. *Main job is teaching* takes the value of 1 if the teacher's main position in the school involves teaching (as opposed to administrative duties). *Enrollment* measures the number of students in the school. *Average SES of students in school* is a 1-4 index measuring the average socioeconomic status of the school's students that participated in SIMCE in that year, and *School average SIMCE score* is the raw average score that students got in that test the year before the teacher was evaluated. *School won previous SNED* is a dummy taking the value of 1 if the school won the previous edition of SNED. *Refused evaluation before* is a dummy variable taking the value of 1 if the teacher had refused to be evaluated before, and *Reevaluated* is a dummy for whether the teacher was reevaluated by 2015.

Table 3: Effect of Assignment to Remediation on Reevaluation Results

| | Overall Score | > Basic | Portfolio | Peer | Supervisor | Self |
|---|---|---|---|---|---|---|
| **Panel A: Without Year FE and Controls** | | | | | | |
| Basic | 0.026** | 0.050** | 0.045*** | 0.031 | -0.048 | -0.002 |
| | (0.013) | (0.021) | (0.013) | (0.033) | (0.035) | (0.017) |
| Bandwidth | 0.192 | 0.208 | 0.234 | 0.205 | 0.207 | 0.204 |
| Number of Teachers | 16,774 | 17,983 | 19,972 | 17,854 | 17,956 | 17,646 |
| Dependent Var. Mean | 2.630 | 0.711 | 2.233 | 3.072 | 2.915 | 3.877 |
| **Panel B: Including Year FE and Controls** | | | | | | |
| Basic | 0.026** | 0.050** | 0.045*** | 0.024 | -0.050 | -0.007 |
| | (0.013) | (0.021) | (0.013) | (0.032) | (0.034) | (0.017) |
| Bandwidth | 0.198 | 0.224 | 0.233 | 0.210 | 0.217 | 0.214 |
| Number of Teachers | 16,511 | 18,453 | 19,015 | 17,347 | 17,974 | 17,658 |
| Dependent Var. Mean | 2.630 | 0.712 | 2.231 | 3.074 | 2.916 | 3.877 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The table presents the results of a fuzzy RD using a local linear regression, where the dependent variable is specified in the column header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and *> Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self assessment, respectively. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. The regressions in Panel B include year fixed effects and teacher and school controls measured at the year of the teachers' first evaluation (in the school where they taught during that year). The teacher and school characteristics are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 4: Effect of Assignment to Remediation on Students' Standardized Test Scores

| | Number of Years After First Evaluation ($t_0$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Year $1^{st}$ Ev. ($t_0$) | $t_0$+1 | $t_0$+2 | $t_0$+3 | Year $2^{nd}$ Ev. ($t_0$+4) | $t_0$+5 | $t_0$+6 |
| **Panel A: Without Year FE and Controls** | | | | | | | |
| Basic | 0.098 | -0.020 | 0.007 | -0.048 | -0.117** | 0.111* | 0.048 |
| | (0.064) | (0.062) | (0.060) | (0.052) | (0.059) | (0.064) | (0.063) |
| Bandwidth | 0.155 | 0.143 | 0.256 | 0.158 | 0.204 | 0.172 | 0.258 |
| Number of Observations | 92,469 | 120,350 | 181,649 | 141,193 | 150,205 | 136,931 | 161,090 |
| Number of Teachers | 2,048 | 2,605 | 3,924 | 3,140 | 3,524 | 3,288 | 4,077 |
| **Panel B: Including Year FE and Controls** | | | | | | | |
| Basic | 0.056 | -0.024 | -0.034 | -0.051 | -0.092** | 0.090 | 0.051 |
| | (0.060) | (0.053) | (0.051) | (0.044) | (0.045) | (0.058) | (0.054) |
| Bandwidth | 0.114 | 0.133 | 0.227 | 0.172 | 0.239 | 0.156 | 0.236 |
| Number of Observations | 65,699 | 108,135 | 159,393 | 146,119 | 162,741 | 119,323 | 143,470 |
| Number of Teachers | 1,467 | 2,356 | 3,437 | 3,233 | 3,792 | 2,858 | 3,621 |

*Notes:* The sample is composed of all 4th, 8th and 10th grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 in math, language, and natural and social science, and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between one and six years after. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable for whether the teacher responsible for that subject obtained a Basic score in his/her first evaluation, and the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year). I employ a fuzzy RD using a local linear regression, without any controls or fixed effects in Panel A, and controlling for year fixed effects, subjectxgrade fixed effects, the baseline characteristics of teachers and their schools, and the SES and gender of students in Panel B. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the student's school are presented in parentheses. Teacher and school specific controls are measured at the year of the teachers' first evaluation (in the school where they taught during that year), and are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. Regressions in Panel B also include dummies for the student's gender and the average SES of students in his/her class. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Effect of Assignment to Remediation on Attrition and Sorting

| | Coefficient | Standard Error | Mean |
|---|---|---|---|
| **Panel A: Reevaluation Rate** | | | |
| Reevaluated | -0.024 | (0.020) | 0.646 |
| Years until reevaluated | 0.003 | (0.043) | 4.292 |
| **Panel B: Working Conditions at Reevaluation** | | | |
| Job Characteristics | | | |
| Changed grade/subject | -0.018 | (0.019) | 0.244 |
| Changed municipality | 0.008 | (0.011) | 0.060 |
| Changed school | -0.018 | (0.020) | 0.240 |
| Civil servant | 0.010 | (0.025) | 0.725 |
| Total contract hours | -0.127 | (0.417) | 38.003 |
| More than one school | 0.006 | (0.016) | 0.119 |
| More than one municipality | 0.001 | (0.012) | 0.085 |
| Evaluated Subject and Level | | | |
| Math | 0.001 | (0.013) | 0.127 |
| Language | 0.011 | (0.014) | 0.131 |
| Social science | 0.001 | (0.011) | 0.089 |
| Natural science | 0.002 | (0.012) | 0.093 |
| English | -0.003 | (0.009) | 0.047 |
| Art or music | -0.014* | (0.008) | 0.042 |
| Physical education | -0.007 | (0.009) | 0.060 |
| Other subjects | 0.005 | (0.011) | 0.069 |
| Lower primary school | 0.003 | (0.022) | 0.326 |
| Upper primary school | -0.007 | (0.024) | 0.433 |
| Secondary school | 0.004 | (0.022) | 0.225 |
| School Characteristics | | | |
| Urban | 0.015 | (0.034) | 0.741 |
| Ln(Enrollment) | -0.037 | (0.097) | 5.738 |
| Teacher-Student Ratio | -0.183 | (0.896) | 27.441 |
| Average SES of students | 0.009 | (0.068) | 1.889 |
| School average SIMCE | 0.357 | (1.540) | 240.923 |
| School won previous SNED | 0.003 | (0.025) | 0.325 |
| **Panel C: Taught a SIMCE Subject/Grade** | | | |
| 1-6 years after evaluation | 0.009 | (0.021) | 0.570 |
| 1 year after evaluation | 0.018 | (0.018) | 0.256 |
| 2 years after evaluation | -0.014 | (0.016) | 0.187 |
| 3 years after evaluation | 0.001 | (0.016) | 0.229 |
| 4 years after evaluation | 0.008 | (0.015) | 0.168 |
| 5 years after evaluation | -0.004 | (0.015) | 0.207 |
| 6 years after evaluation | -0.006 | (0.015) | 0.167 |
| Number of years (1-6) | 0.003 | (0.057) | 1.213 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. Panels B further restricts the sample to teachers who were also reevaluated by 2015. Every row shows the result of a separate fuzzy RD using a local linear regression, where the dependent variable is the covariate specified in the row header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. The first column presents the fuzzy RD bias-corrected coefficient, and the second presents the robust bias-corrected standard errors, clustered by the teacher's municipality. The third column presents the mean value of the dependent variable. *Reevaluated* is a dummy taking the value of 1 if the teacher was reevaluated by 2015, and *Years until reevaluated* is the number of years between the first and second evaluation, for the sample of teachers who were reevaluated. *1-6 years after evaluation* is a dummy equal to one if the teacher taught a subject and grade that participated in Chile's standardized tests at some point in the six years after his/her first evaluation. Analogously, there is a dummy for whether the teacher taught one of these grades and subjects in each of the six years after the evaluation, and *Number of years (1-6)* measures the number of years in which the teacher taught one of these grades and subjects. All the outcomes in Panel B are measured at the year of the teachers' second evaluation, in the school in which they were evaluated, except for the school's average SIMCE score which is lagged by one year. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Effect of Assignment to Remediation on Actions for Improvement and Perceptions

| | Measures After First Evaluation | | | | Improved After First Evaluation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Improved Weaknesses | Asked for Support | Discussed with Colleagues | Met with Principal | Prestige | Job Satisfaction | Job Stability | Responsibilities | Income | Professional Development |
| **Panel A: Without Year FE and Controls** | | | | | | | | | | |
| Basic | -0.021 | 0.089*** | 0.033 | 0.037* | -0.312*** | -0.306*** | -0.107*** | -0.070*** | -0.313*** | 0.085*** |
| | (0.021) | (0.017) | (0.022) | (0.021) | (0.024) | (0.027) | (0.017) | (0.022) | (0.020) | (0.024) |
| Bandwidth | 0.224 | 0.236 | 0.233 | 0.211 | 0.201 | 0.205 | 0.258 | 0.232 | 0.230 | 0.250 |
| Number of Teachers | 17,923 | 18,777 | 18,581 | 11,142 | 12,201 | 12,546 | 15,293 | 13,980 | 13,886 | 14,813 |
| Dependent Var. Mean | 0.625 | 0.178 | 0.369 | 0.131 | 0.294 | 0.577 | 0.155 | 0.264 | 0.268 | 0.311 |
| **Panel B: Including Year FE and Controls** | | | | | | | | | | |
| Basic | -0.023 | 0.086*** | 0.030 | 0.047** | -0.308*** | -0.307*** | -0.099*** | -0.062*** | -0.314*** | 0.083 |
| | (0.021) | (0.018) | (0.020) | (0.022) | (0.025) | (0.027) | (0.018) | (0.022) | (0.020) | (0.023) |
| Bandwidth | 0.236 | 0.227 | 0.264 | 0.215 | 0.192 | 0.196 | 0.245 | 0.237 | 0.243 | 0.261 |
| Number of Teachers | 17,976 | 17,355 | 19,738 | 11,734 | 11,250 | 11,532 | 14,036 | 13,574 | 13,882 | 14,722 |
| Dependent Var. Mean | 0.627 | 0.178 | 0.372 | 0.131 | 0.293 | 0.577 | 0.153 | 0.262 | 0.270 | 0.310 |
| Response Rate | 0.970 | 0.970 | 0.970 | 0.981 | 0.906 | 0.908 | 0.907 | 0.908 | 0.908 | 0.903 |
| P-Value (equality in response rates) | 0.587 | 0.587 | 0.587 | 0.618 | 0.941 | 0.211 | 0.566 | 0.733 | 0.926 | 0.523 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The data is taken from teachers' responses in a survey conducted after the reevaluation process, but before the evaluation results are announced. The table presents the results of a fuzzy RD using a local linear regression, where the running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The dependent variable in columns 1 to 4 is a dummy taking the value of 1 if after the first evaluation, the teacher took measures to improve his/her weaknesses, asked for support to interpret the results of the evaluation, discussed the results with colleagues, or met the principal to discuss the results, respectively. The dependent variable in columns 5 to 10 is a dummy taking the value of 1 if the teacher felt that as a consequence of his/her first evaluation, there was an improvement in the dimension specified in the column header. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. The regressions in Panel B include year fixed effects and teacher and school controls measured at the year of the teachers' first evaluation (in the school where they taught during that year). The teacher and school characteristics are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. The variables in columns 1-3 are only available for teachers who were reevaluated in 2009-2015, the one in column 4 for teachers in 2011-2015, and those in columns 5-10 for those who were reevaluated in 2010-2014. Average response rates for teachers reevaluated in the corresponding period are specified, as well as the p-value of a fuzzy RD regression testing whether there are differential response rates for individuals with a Basic and Competent score in their first evaluation. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Effect of Assignment to Remediation on Reevaluation Results – Heterogeneous Effects by Average Attendance to Remediation

| | Overall Score | > Basic | Portfolio | Peer | Supervisor | Self |
|---|---|---|---|---|---|---|
| **Panel A: Remedial Training Attendance in Municipality Below the Median** | | | | | | |
| Basic | 0.028 | 0.061** | 0.034 | 0.049 | -0.004 | -0.025 |
| | (0.020) | (0.031) | (0.022) | (0.051) | (0.047) | (0.023) |
| Bandwidth | 0.233 | 0.262 | 0.242 | 0.231 | 0.245 | 0.256 |
| Number of Teachers | 7,498 | 8,313 | 7,779 | 7,452 | 7,844 | 8,155 |
| Dependent Var. Mean | 2.623 | 0.703 | 2.226 | 3.087 | 2.851 | 3.875 |
| **Panel B: Remedial Training Attendance in Municipality Above the Median** | | | | | | |
| Basic | 0.030 | 0.067** | 0.048** | 0.044 | -0.077 | 0.020 |
| | (0.019) | (0.031) | (0.024) | (0.057) | (0.050) | (0.023) |
| Bandwidth | 0.260 | 0.263 | 0.211 | 0.180 | 0.251 | 0.249 |
| Number of Teachers | 8,352 | 8,422 | 6,876 | 5,991 | 8,011 | 7,986 |
| Dependent Var. Mean | 2.648 | 0.732 | 2.253 | 3.099 | 2.876 | 3.879 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated between 2010 and 2015. Panel A (B) presents the results for the sample of teachers who work in a municipality in which the average attendance to remedial training of reevaluated teachers with a Basic rating was below (above) the median. The table presents the results of a fuzzy RD using a local linear regression, where the dependent variable is specified in the column header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and $>$ *Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self assessment, respectively. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 8: Effect of Assignment to Remediation on Reevaluation Results – Heterogeneous Effects by Share of Peers Receiving AVDI

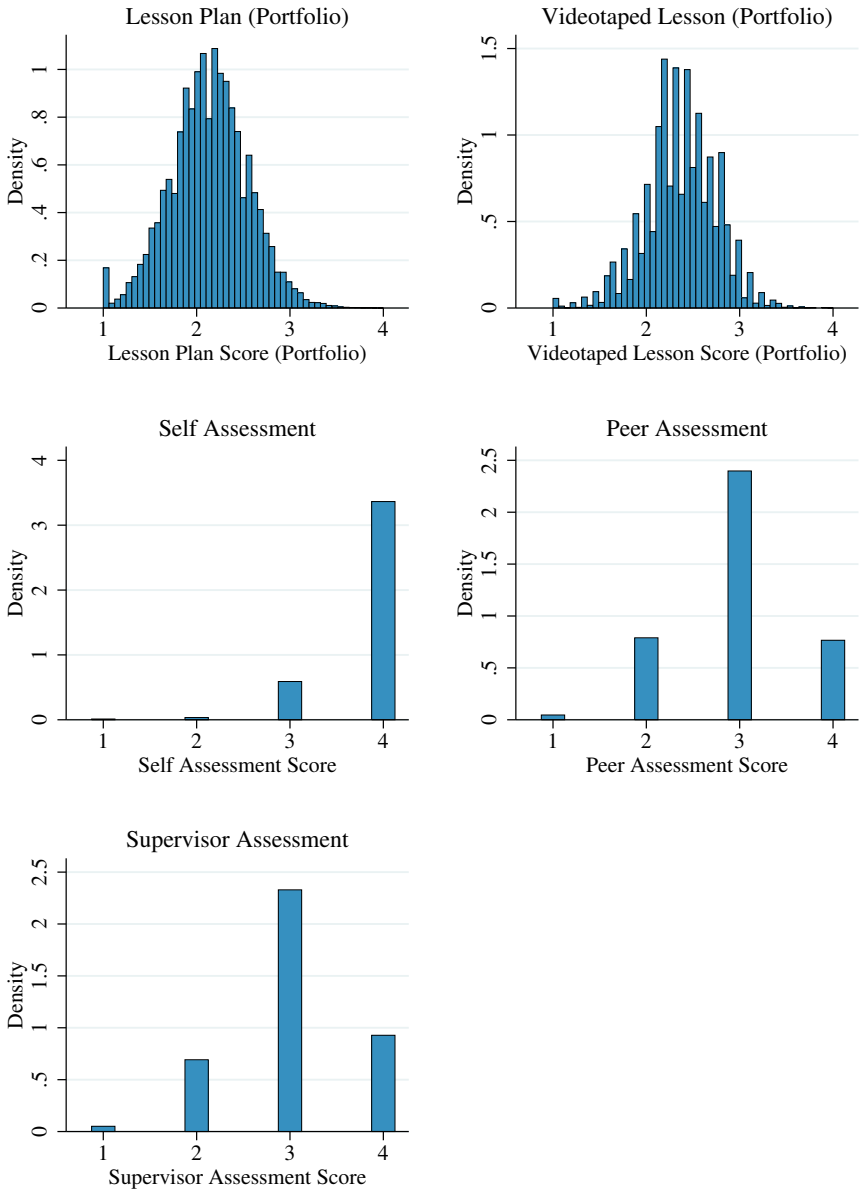
| | Overall Score | $>$ Basic | Portfolio | Peer | Supervisor | Self |
|---|---|---|---|---|---|---|
| **Panel A: Share of AVDI Winners in School Below the Median** | | | | | | |
| Basic | 0.027* | 0.041 | 0.045** | 0.022 | -0.030 | -0.021 |
| | (0.015) | (0.029) | (0.019) | (0.039) | (0.047) | (0.022) |
| Bandwidth | 0.255 | 0.235 | 0.228 | 0.243 | 0.238 | 0.239 |
| Number of Teachers | 9,480 | 8,905 | 8,723 | 9,120 | 8,979 | 9,007 |
| Dependent Var. Mean | 2.615 | 0.688 | 2.217 | 3.055 | 2.882 | 3.872 |
| **Panel B: Share of AVDI Winners in School Above the Median** | | | | | | |
| Basic | 0.033* | 0.059** | 0.050*** | 0.037 | -0.050 | 0.023 |
| | (0.017) | (0.025) | (0.017) | (0.041) | (0.043) | (0.021) |
| Bandwidth | 0.228 | 0.281 | 0.234 | 0.238 | 0.223 | 0.199 |
| Number of Teachers | 11,820 | 14,092 | 12,066 | 12,227 | 11,677 | 10,624 |
| Dependent Var. Mean | 2.647 | 0.737 | 2.245 | 3.090 | 2.950 | 3.879 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. Panel A (B) presents the results for the sample of teachers who work in a school in which the share of teachers receiving the AVDI award was below (above) the median at the moment of their reevaluation. The sample was split within a given year and municipality. The table presents the results of a fuzzy RD using a local linear regression, where the dependent variable is specified in the column header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and $>$ *Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self assessment, respectively. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

# ONLINE APPENDIX

## Appendix A Appendix Figures and Tables

Figure A.1: Distribution of Subitem Scores for Teachers First Evaluated in 2004-2010




*Notes*: The sample includes all public school teachers instructing primary and secondary school who were evaluated for the first time in 2004-2010. Each figure plots the distribution of scores for the evaluation subitem specified in the column header.

Figure A.2: Example of Feedback Report to Teachers

**Dimension F: Classroom learning environment**

The information to rate this dimension was obtained from the videotaped lesson.

Concerning this dimension, a competent teacher is characterized by:

- achieving that his/her students remain focused on the activities, maintaining a code of conduct that favors the development of the lecture,
- generating equal opportunities for students' participation and promoting the interaction between them, and
- delivering clear instructions and adequately monitoring the development of the class activities.

**Description of the aspects that were achieved**

Manages to keep his/her students focused on the activities and respecting the class rules. This generates a work environment that is conducive towards learning.

Adequately accompanies and monitors the activities carried out by the students, giving them clear and precise instructions to orient their work.

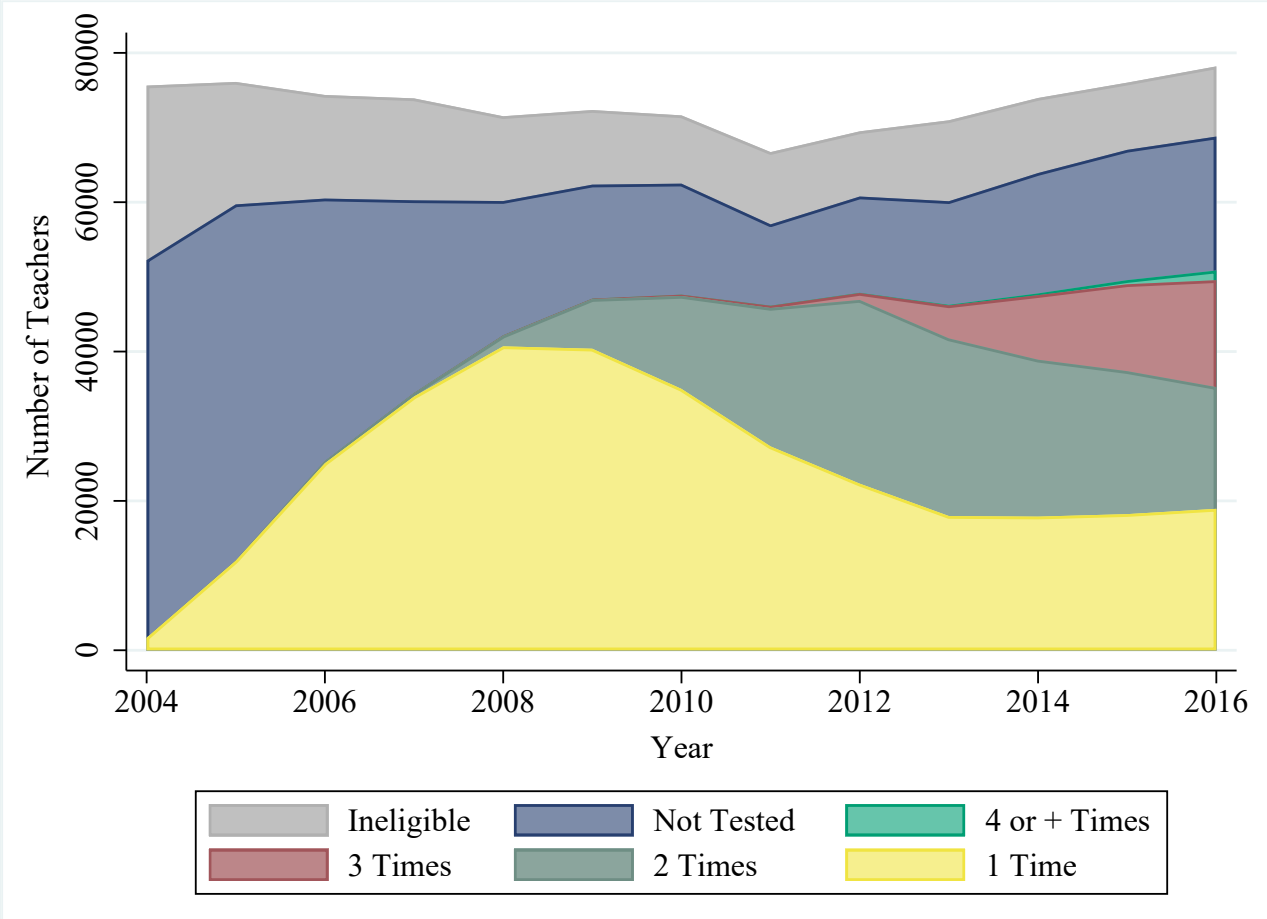
**Description of the aspects that need to be developed**

Although the teacher manages to make the students participate in the different stages of the lecture in an equitable manner, his/her performance has certain weaknesses, since he/she does not promote the interaction or collaboration between students.

Considering the aforementioned aspects, [teacher's name] obtains a **Competent achievement level** in dimension F, *Classroom learning environment*.

*Notes*: Self-translation based on the example of the 2009 reports available at http://www.docentemas.cl/docs/ejem_eval_ind.pdf.

Figure A.3: Rollout and Compliance with Teacher Evaluations in 2004-2016




*Notes*: The sample is composed of all public primary and secondary school teachers in Chile in 2004-2016. Teachers who are three or less years away from the retirement age (65 for men, 60 for women) can opt out, and those who are in their first year in the public school system are not eligible for evaluation. The pool of teachers in each year changes as some teachers exit the public school system and others enter it.

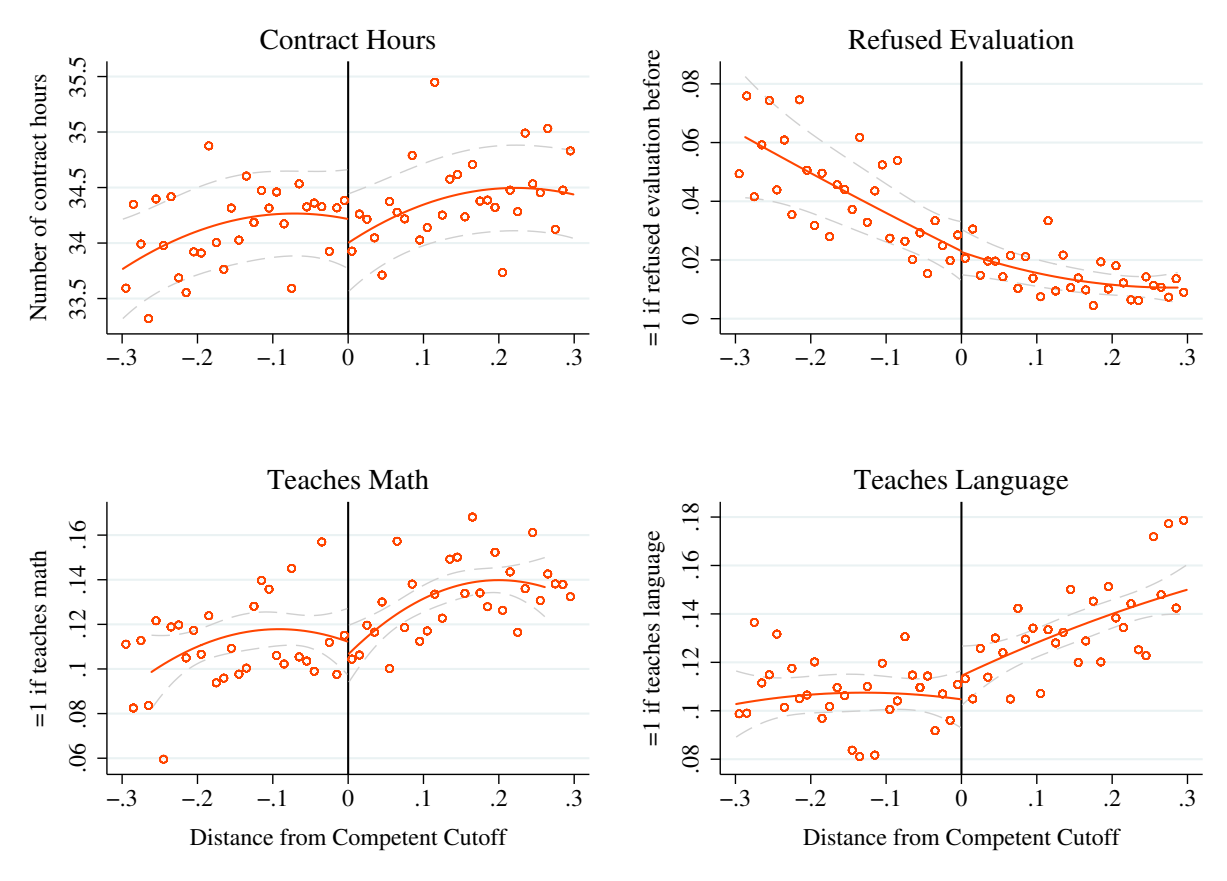
Figure A.4: Balance of Covariates in Full Sample of Teachers


*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010, for which information on teacher characteristics was available (96% of those first evaluated in this period). The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. *Male* is a dummy for whether the is male, *Age* and *Years of Experience* measure the teacher's age and number of years as a teacher at the moment of the evaluation. *Has a Degree* is a dummy for whether the teacher holds a degree.

Figure A.5: Balance of Covariates in Full Sample of Teachers

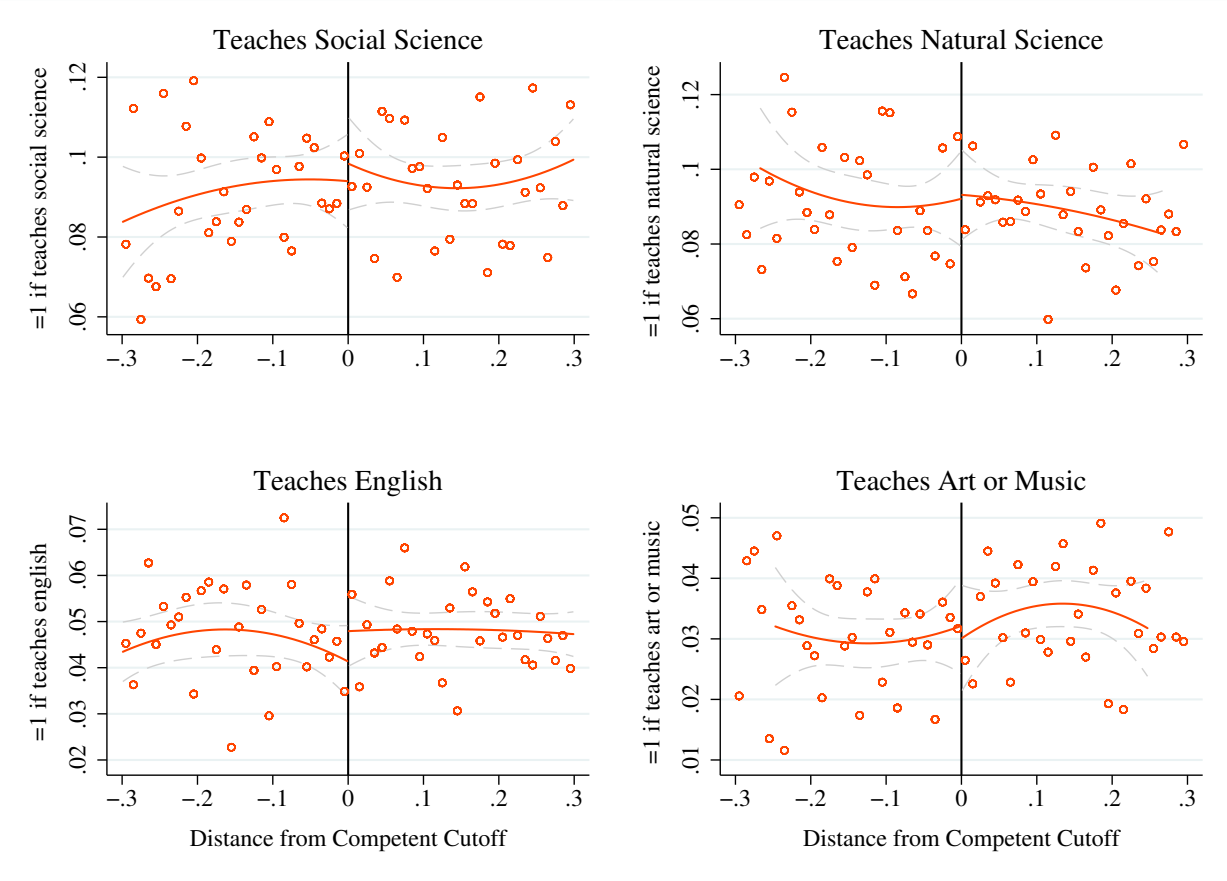

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010, for which information on teacher characteristics was available (96% of those first evaluated in this period). The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. *Main Job is Teaching* takes a value of 1 if the teacher's main position in the school involves teaching (as opposed to administrative duties), and *Works in More Than One School* and *Works in More Than One Municipality* are dummies for whether the teacher worked in more than one school or municipality in the year of the first evaluation. *Civil Servant* is a dummy variable taking the value of 1 if the teacher has a civil servant position, and 0 if he/she is a contract teacher.

Figure A.6: Balance of Covariates in Full Sample of Teachers



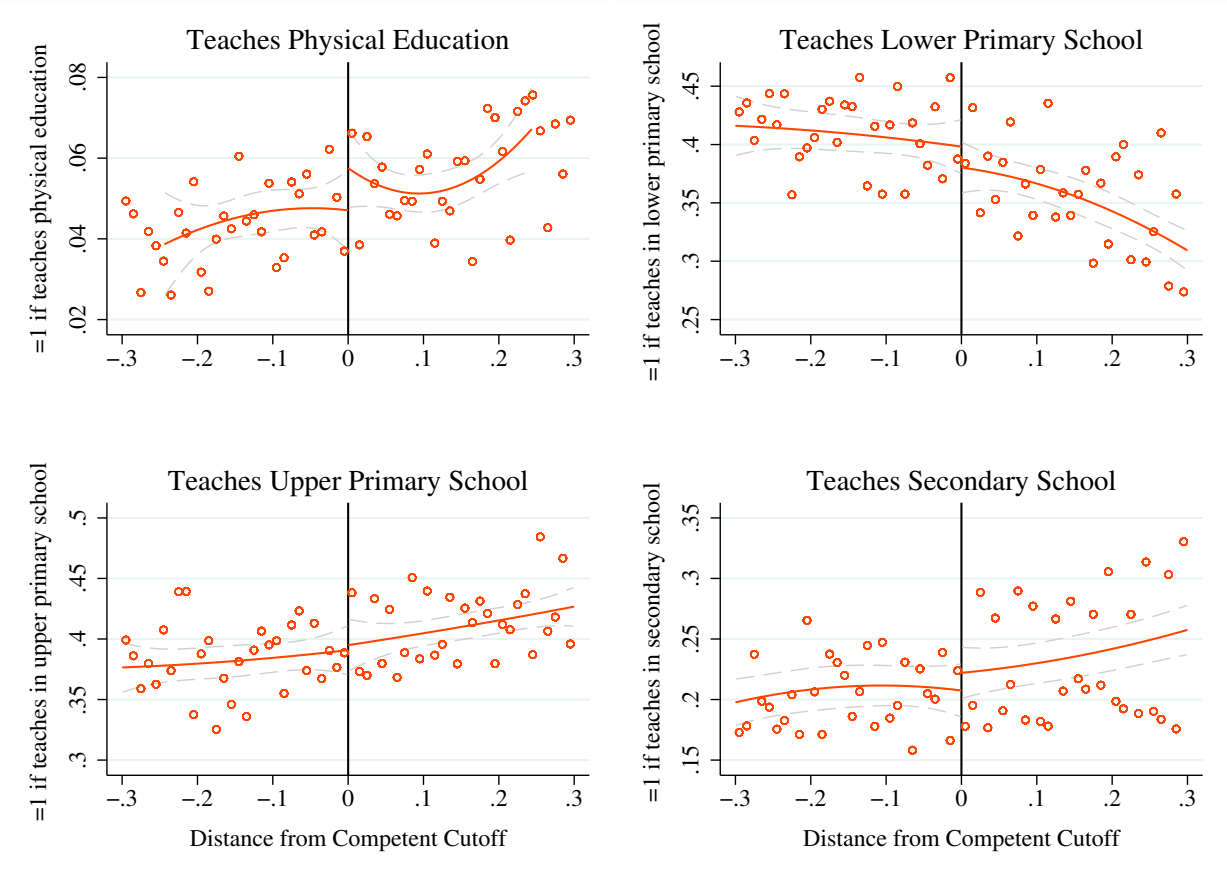

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. The sample in the first graphs is further restricted to teachers for which information on teacher characteristics was available (96% of those first evaluated in this period). The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. *Contract Hours* are the number of hours a week that teachers are contractually obliged to work in the school in which they were evaluated, and *Refused Evaluation* is a dummy variable taking the value of 1 if the teacher had refused to be evaluated before. *Teaches Math* and *Teaches Language* are dummies for whether the teacher was assessed in math and language in his/her first evaluation.

Figure A.7: Balance of Covariates in Full Sample of Teachers



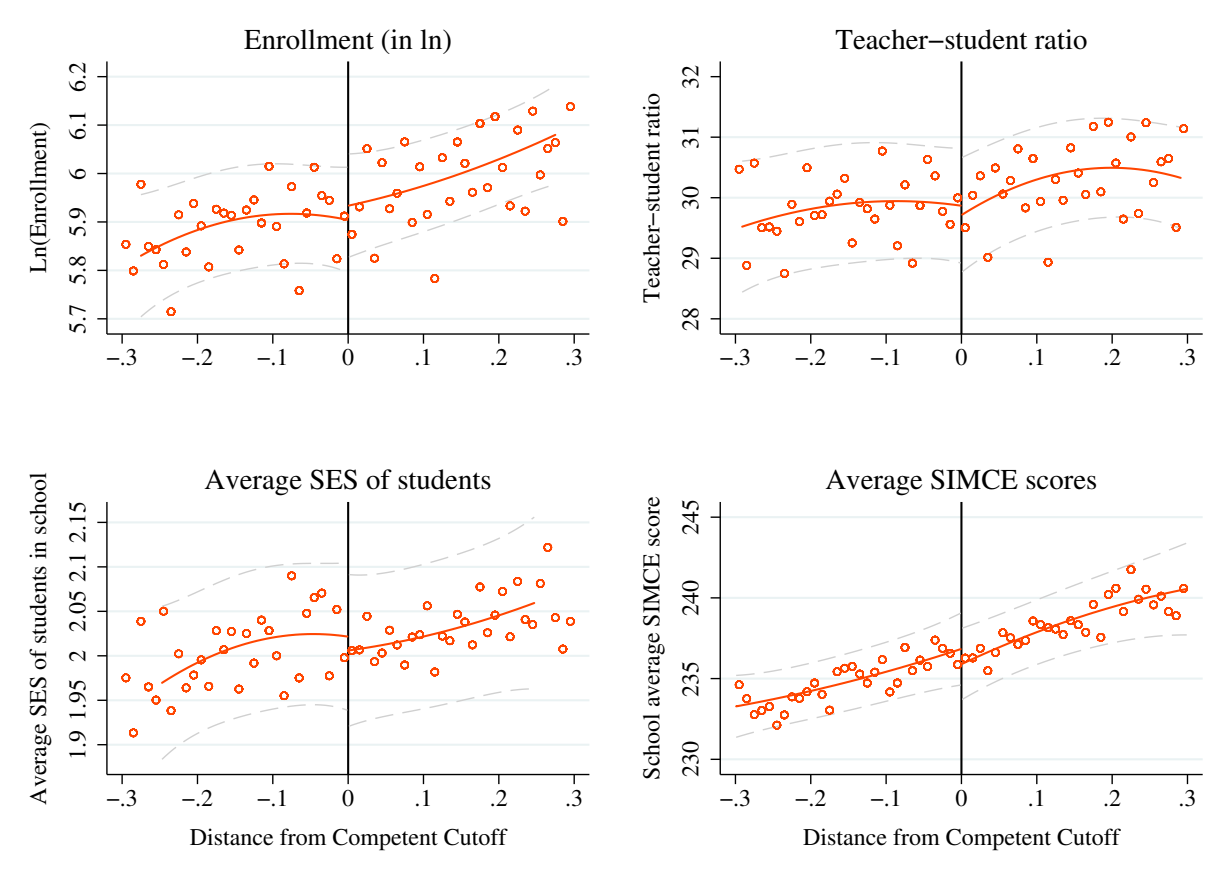

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. The dependent variables are dummies for whether the teacher was assessed in the corresponding subject in his/her first evaluation.

Figure A.8: Balance of Covariates in Full Sample of Teachers



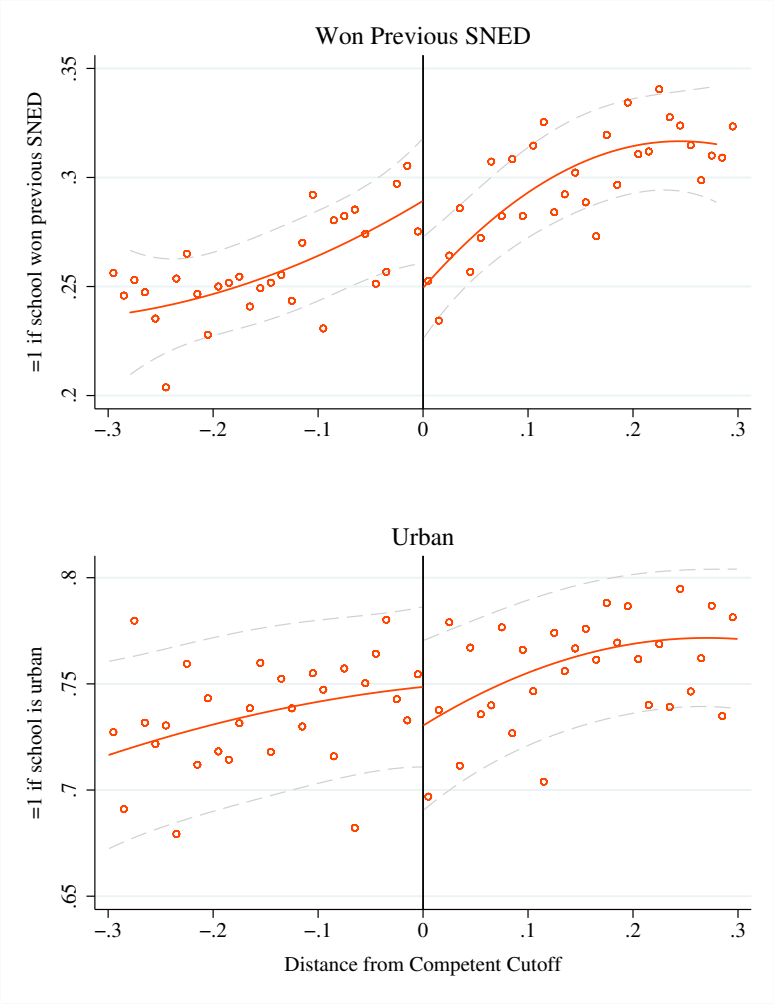

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010. The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. *Teaches Physical Education* is a dummy for whether the teacher was assessed in physical education in his/her first evaluation. *Teaches Lower Primary School*, *Teaches Upper Primary School* and *Teaches Secondary School* are dummies for whether he/she was assessed at the lower primary ($1^{st}$ to $4^{th}$ grade), upper primary ($5^{th}$ to $8^{th}$ grade) or secondary school ($9th^{th}$ to $12^{th}$ grade) level.

Figure A.9: Balance of Covariates in Full Sample of Teachers



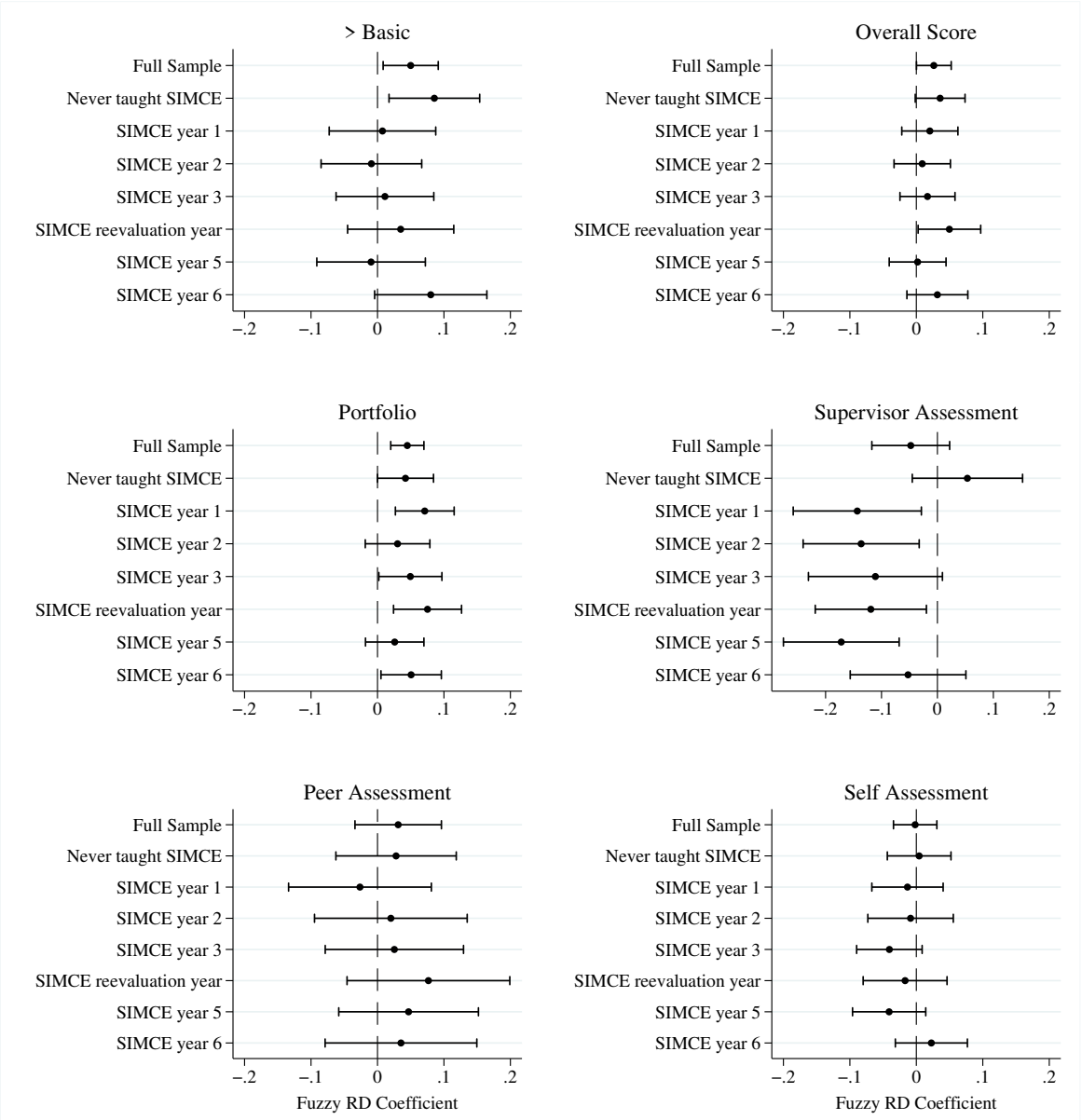

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010, for which a school identifier was available (around 99.6% of those first evaluated in this period). The sample in the bottom right graph is further restricted to teachers who worked in a school large enough for average test scores to be reported (approximately 92% of the sample). The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed grey lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. All of the dependent variables are school level characteristics at the year of the teacher's first evaluation, in the school in which they were evaluated. *Ln(Enrollment)* is the number of students in the school (in ln), and *Teacher-student ratio* is the school's average number of students per teacher. *Average SES of Students* is a 1-4 index measuring the average socioeconomic status of the school's students that participated in SIMCE in that year, and *Average SIMCE Scores* is the raw average score that the school's students got in that test the year before.

Figure A.10: Balance of Covariates in Full Sample of Teachers



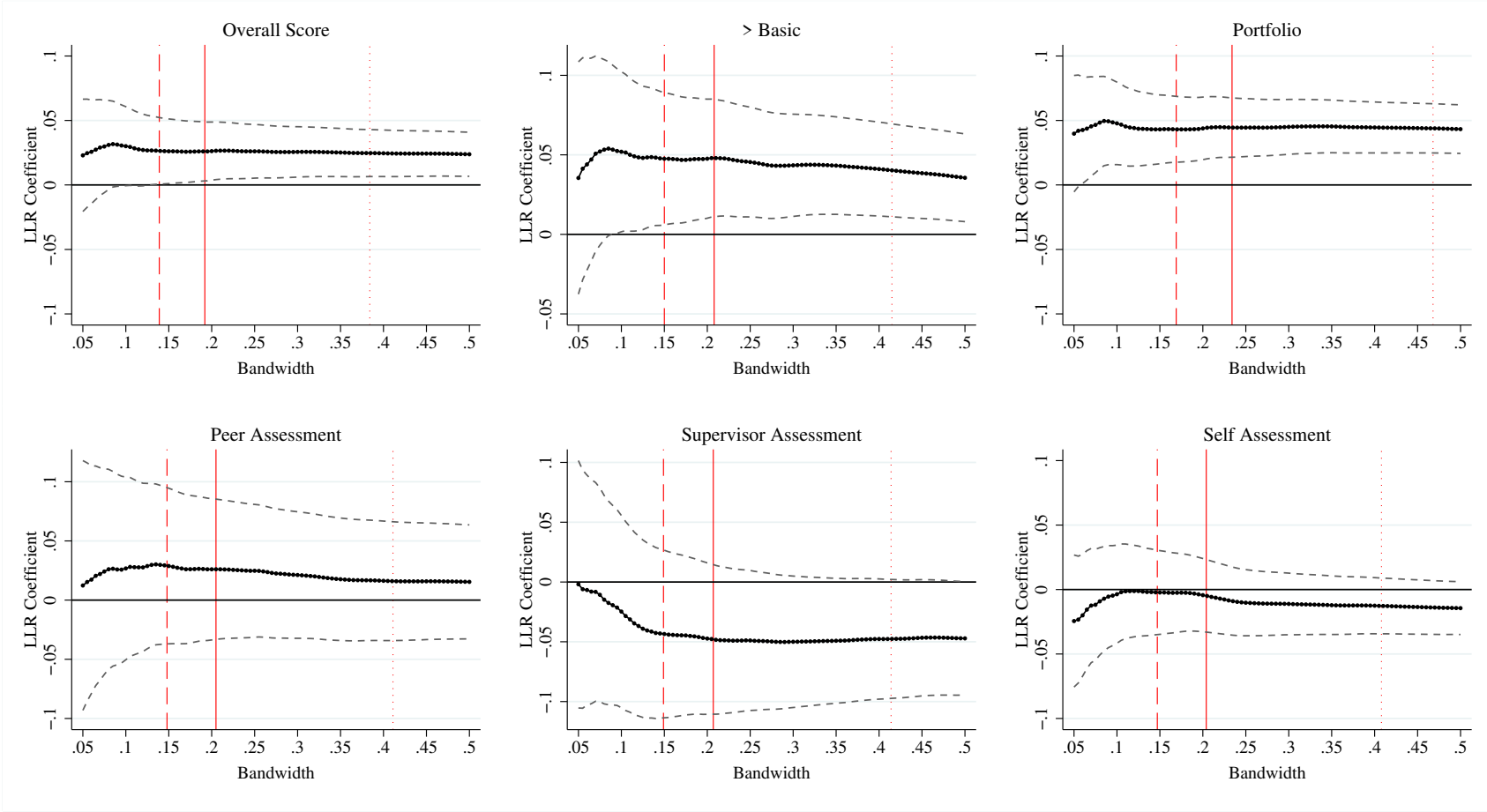

*Notes*: The sample is composed of all public primary and secondary school teachers in Chile evaluated for the first time in 2004-2010, for which a school identifier was available (around 99.6% of those first evaluated in this period) The running variable is the evaluation score centered around the Basic/Competent threshold. Dots represent bin averages for a bin width of 0.01. The solid orange line plots the fitted values of a quadratic regression over the MSE-optimal bandwidth, and the dashed gray lines are the 95% robust bias-corrected confidence intervals with standard errors clustered by the teacher's municipality. All of the dependent variables are school level characteristics at the year of the teacher's first evaluation, in the school in which they were evaluated. *Won Previous SNED* is a dummy taking the value of 1 if the school won the previous edition of SNED (a nationwide teacher pay for performance program), and *Urban* is a dummy for whether the school is located in an urban area.

Figure A.11: Effect of Assignment to Remediation on Reevaluation Results – Subsample of Teachers by SIMCE Participation



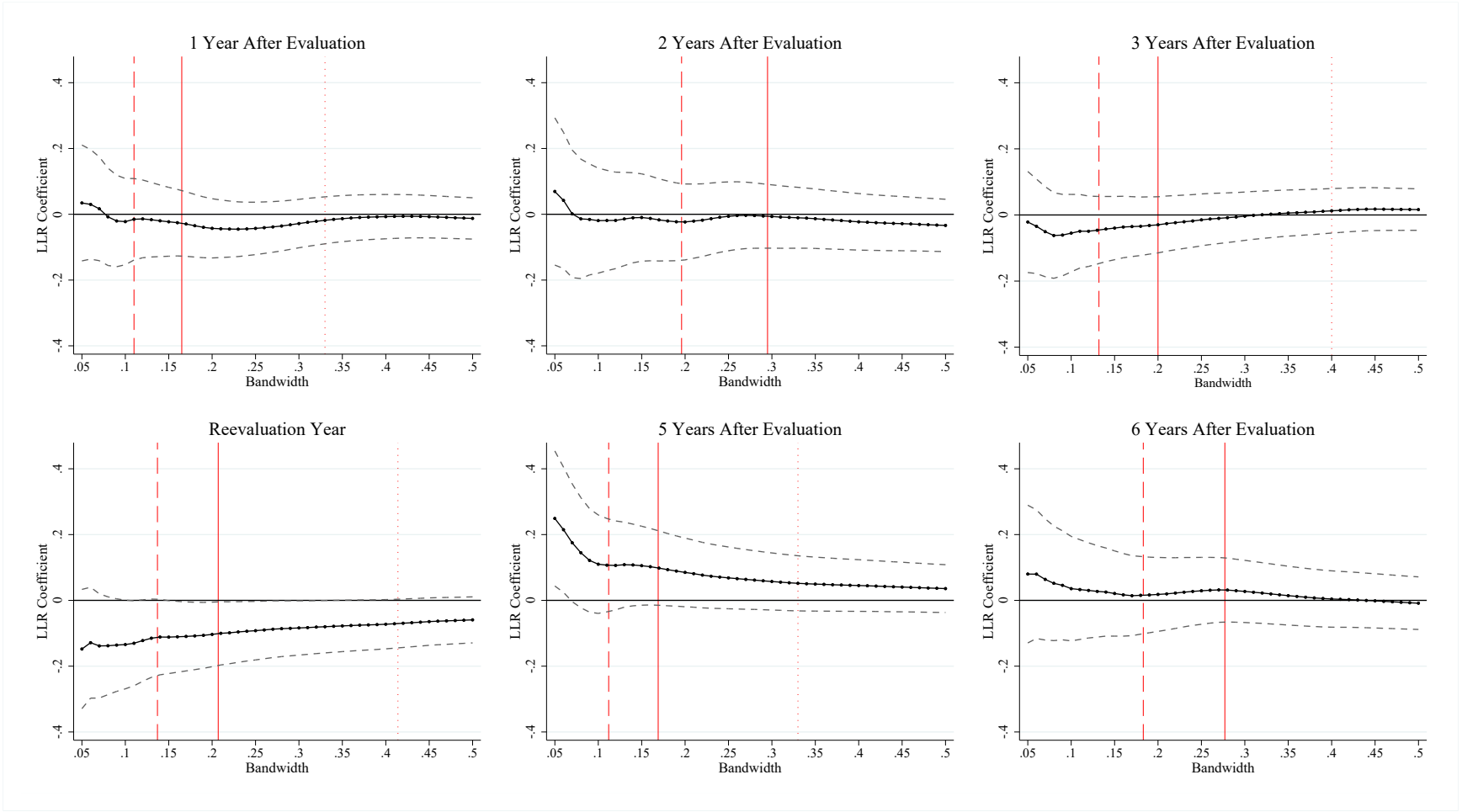

*Notes*: The figures plot the robust bias-corrected coefficients and 95% robust confidence interval of a fuzzy RD using a local linear regression, where the dependent variable is specified in the figure header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. All regressions are run over the MSE-optimal bandwidth, and standard errors are clustered by the teacher's municipality. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and $>$ *Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self assessment, respectively. The labels in the vertical axis indicate the sample over which the regressions were conducted. *Full Sample* includes all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. *Never taught SIMCE* restrict the full sample to teachers who do not instruct a subject or grade that participates in SIMCE any of the six years after their first evaluation. *SIMCE year 1* restricts the full sample to teachers who instruct a subject/grade that participates in SIMCE the year after their first evaluation. The analogous definition applies to the remaining subsamples.

Figure A.12: Effect of Assignment to Remediation on Reevaluation Results – Bandwidth Sensitivity




*Notes*: The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The figures plot the point estimates and 95% conventional confidence interval of a fuzzy RD using a local linear regression, where the dependent variable is specified in the figure header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. I re-estimate the RD coefficient for each bandwidth between 0.05 and 0.5, in 0.05 increments. The dashed line marks the CER optimal bandwidth, and the solid and dotted lines indicate the MSE optimal bandwidth and two times its value. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and $>$ *Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self assessment, respectively. Standard errors are clustered by the teacher's municipality.

Figure A.13: Effect of Assignment to Remediation on Students' Standardized Test Scores – Bandwidth Sensitivity




*Notes*: The sample is composed of all $4^{th}$, $8^{th}$ and $10^{th}$ grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 in math, language, and natural and social science, and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between 1 and 6 years after. The figures plot the point estimates and 95% conventional confidence interval of a fuzzy RD using a local linear regression, where the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year), and the running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. I re-estimate the RD coefficient for each bandwidth between 0.05 and 0.5, in 0.1 increments. The dashed line marks the CER optimal bandwidth, and the solid and dotted lines indicate the MSE optimal bandwidth and two times its value. Standard errors are clustered by the student's school.

Table A.1: Characteristics of Remedial Training Courses

| | Mean | SD | Response Rate | Survey Years |
|---|---|---|---|---|
| Participated in remediation | 0.79 | 0.40 | 0.92 | 2010-2015 |
| Remediation will improve evaluation scores | 0.67 | 0.47 | 0.93 | 2013-2015 |
| **Topics covered in remediation** | | | | |
| Lesson planning | 0.71 | 0.46 | 1.00 | 2011-2015 |
| Student evaluations | 0.67 | 0.47 | 1.00 | 2011-2015 |
| Learning environment | 0.35 | 0.48 | 1.00 | 2011-2015 |
| Content knowledge | 0.20 | 0.40 | 1.00 | 2011-2015 |
| Pedagogical skills | 0.28 | 0.45 | 1.00 | 2011-2015 |
| Reflection about own teaching practices | 0.53 | 0.50 | 1.00 | 2011-2015 |
| **Format of remediation** | | | | |
| Lectures | 0.67 | 0.47 | 0.68 | 2012-2015 |
| Group discussion | 0.19 | 0.39 | 0.68 | 2012-2015 |
| Reading list | 0.04 | 0.19 | 0.68 | 2012-2015 |
| Classroom observation | 0.01 | 0.08 | 0.68 | 2012-2015 |
| Role-play or simulation | 0.01 | 0.11 | 0.68 | 2012-2015 |
| Mentoring or coaching | 0.04 | 0.19 | 0.68 | 2012-2015 |
| Development of teacher networks by subject | 0.02 | 0.12 | 0.68 | 2012-2015 |
| Analysis of videotaped lessons | 0.01 | 0.10 | 0.68 | 2012-2015 |
| Online courses | 0.01 | 0.11 | 0.68 | 2013-2015 |
| **Quality of remediation (1-7)** | | | | |
| Quality of the activities | 4.98 | 1.64 | 0.97 | 2010-2015 |
| Relevant to teacher's weaknesses | 5.01 | 1.69 | 0.97 | 2010-2015 |
| Useful | 5.09 | 1.69 | 0.97 | 2011-2015 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010, obtained a Basic rating, and were reevaluated between 2008 and 2015. The data is taken from teachers' responses in a survey conducted after the reevaluation process, but before the evaluation results are announced. The last column indicates the period in which the corresponding measure was included in the teacher survey, and the third column indicates the percentage of teachers who responded the question, out of those who took the reevaluation in a year in which the question was asked. Except for the first row, all other statistics are only available for teachers who declared to participate in remedial training (including response rates).

| | | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2nd | Language | | | | | | | | | X | X | X | X | |
| 4th | Language | | X | X | X | X | X | X | X | X | X | X | X | X |
| | Math | | X | X | X | X | X | X | X | X | X | X | X | X |
| | Natural Sciences | | | | | | X | | X | | X | | | |
| | Social Sciences | | | | | X | | X | | X | | X | | |
| 6th | Language | | | | | | | | | | X | X | X | X |
| | Math | | | | | | | | | | X | X | X | X |
| | Natural Sciences | | | | | | | | | | | X | | |
| | Social Sciences | | | | | | | | | | | | X | X |
| 8th | Language | X | | | X | | X | | X | | X | X | X | |
| | Math | X | | | X | | X | | X | | X | X | X | |
| | Natural Sciences | X | | | X | | X | | X | | X | | X | |
| | Social Sciences | X | | | X | | X | | X | | | X | | |
| 10th | Language | | | X | | X | | X | | X | X | X | X | X |
| | Math | | | X | | X | | X | | X | X | X | X | X |
| | Natural Sciences | | | | | | | | | | | X | | X |
| | Social Sciences | | | | | | | | | | | | X | |

Table A.3: Effect of Assignment to Remediation on Portfolio Reevaluation Results

| | Pedagogical Unit Plan | | | | Videotaped Lesson | | |
|---|---|---|---|---|---|---|---|
| | Organization | Lesson Analysis | Ev. Quality | Ev. Analysis | Learning Env. | Structure | Interactions |
| **Panel A: Without Year FE and Controls** | | | | | | | |
| Basic | 0.062** | 0.067** | 0.027 | 0.117*** | 0.005 | 0.002 | -0.004 |
| | (0.025) | (0.026) | (0.029) | (0.029) | (0.013) | (0.023) | (0.022) |
| Bandwidth | 0.222 | 0.202 | 0.206 | 0.201 | 0.190 | 0.195 | 0.210 |
| Number of Teachers | 19,056 | 17,535 | 17,857 | 17,535 | 16,638 | 17,121 | 18,090 |
| Dependent Var. Mean | 2.346 | 2.130 | 2.093 | 2.040 | 2.723 | 2.378 | 1.989 |
| **Panel B: Including Year FE and Controls** | | | | | | | |
| Basic | 0.066*** | 0.071*** | 0.024 | 0.111*** | 0.008 | 0.005 | 0.001 |
| | (0.025) | (0.023) | (0.029) | (0.027) | (0.013) | (0.021) | (0.019) |
| Bandwidth | 0.230 | 0.231 | 0.213 | 0.218 | 0.191 | 0.210 | 0.226 |
| Number of Teachers | 18,904 | 18,904 | 17,565 | 18,002 | 15,931 | 17,329 | 18,544 |
| Dependent Var. Mean | 2.343 | 2.131 | 2.091 | 2.039 | 2.721 | 2.373 | 1.988 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The table presents the results of a fuzzy RD using a local linear regression, where the dependent variable is specified in the column header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The dependent variables in columns (1) to (4) are subitems of the portion of the portfolio devoted to planning and implementing a pedagogical unit, whereas those in columns (5)-(7) are components of the videotaped lessons. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. The regressions in Panel B include year fixed effects and teacher and school controls measured at the year of the teachers' first evaluation (in the school where they taught during that year). The teacher and school characteristics are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.4: Effect of Assignment to Remediation on Students' Standardized Test Scores

| | Number of Years After First Evaluation ($t_0$) | | | |
|---|---|---|---|---|
| | Year $1^{st}$ Ev. ($t_0$) | $t_0$+1 to $t_0$+3 | Year $2^{nd}$ Ev. ($t_0$+4) | $t_0$+5 to $t_0$+6 |
| **Panel A: Without Year FE and Controls** | | | | |
| Basic | 0.098 | -0.032 | -0.117** | 0.068 |
| | (0.064) | (0.041) | (0.059) | (0.050) |
| Bandwidth | 0.155 | 0.153 | 0.204 | 0.196 |
| Number of Observations | 92,469 | 377,547 | 150,205 | 281,521 |
| Number of Teachers | 2,048 | 5,741 | 3,524 | 5,693 |
| **Panel B: Including Year FE and Controls** | | | | |
| Basic | 0.056 | -0.047 | -0.092** | 0.066 |
| | (0.060) | (0.032) | (0.045) | (0.043) |
| Bandwidth | 0.114 | 0.168 | 0.239 | 0.199 |
| Number of Observations | 65,699 | 414,252 | 162,741 | 273,878 |
| Number of Teachers | 1,469 | 6,001 | 3,794 | 5,536 |

*Notes:* The sample is composed of all 4th, 8th and 10th grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 in math, language, and natural and social science, and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between one and six years after. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable for whether the teacher responsible for that subject obtained a Basic score in his/her first evaluation, and the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year). I employ a fuzzy RD using a local linear regression, without any controls or fixed effects in Panel A, and controlling for year fixed effects, subjectxgrade fixed effects, the baseline characteristics of teachers and their schools, and the SES and gender of students in Panel B. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the student's school are presented in parentheses. Teacher and school specific controls are measured at the year of the teachers' first evaluation (in the school where they taught during that year), and are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. Regressions in Panel B also include dummies for the student's gender and the average SES of students in his/her class. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.5: Effect of Assignment to Remediation on Students' Standardized Test Scores – $4^{\text{th}}$ Grade Math and Language

| | Number of Years After First Evaluation ($t_0$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Year $1^{st}$ Ev. ($t_0$) | $t_0$+1 | $t_0$+2 | $t_0$+3 | Year $2^{nd}$ Ev. ($t_0$+4) | $t_0$+5 | $t_0$+6 |
| **Panel A: Mathematics** | | | | | | | |
| Basic | 0.131 | 0.092 | -0.027 | -0.062 | -0.121 | 0.162 | 0.060 |
| | (0.086) | (0.090) | (0.103) | (0.088) | (0.081) | (0.125) | (0.109) |
| Bandwidth | 0.140 | 0.155 | 0.201 | 0.183 | 0.241 | 0.164 | 0.213 |
| Number of Observations | 20,388 | 23,796 | 29,816 | 29,026 | 31,635 | 20,558 | 25,527 |
| Number of Teachers | 1,164 | 1,492 | 1,937 | 1,847 | 2,325 | 1,594 | 1,999 |
| **Panel B: Language** | | | | | | | |
| Basic | 0.090 | -0.019 | -0.048 | -0.018 | -0.179** | 0.063 | 0.085 |
| | (0.103) | (0.079) | (0.079) | (0.083) | (0.084) | (0.074) | (0.104) |
| Bandwidth | 0.070 | 0.186 | 0.278 | 0.146 | 0.161 | 0.270 | 0.213 |
| Number of Observations | 10,282 | 28,421 | 38,766 | 22,565 | 22,206 | 32,279 | 24,573 |
| Number of Teachers | 597 | 1,744 | 2,485 | 1,454 | 1,622 | 2,393 | 1,969 |

*Notes:* The sample is composed of all $4^{\text{th}}$ grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 in math and language, and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the tested subject, and by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between one and six years after. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable for whether the teacher responsible for that subject obtained a Basic score in his/her first evaluation, and the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by subject and year). In Panel A, the dependent variable are math scores, and in Panel B language. I employ a fuzzy RD using a local linear regression over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the student's school score are presented in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.6: Balance in Student Characteristics for Teachers Evaluated at the Basic vs. Competent Level in 2004-2010

| | Number of Years After First Evaluation ($t_0$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Year $1^{st}$ Ev. ($t_0$) | $t_0$+1 | $t_0$+2 | $t_0$+3 | Year $2^{nd}$ Ev. ($t_0$+4) | $t_0$+5 | $t_0$+6 |
| **Panel A: Lagged GPA** | | | | | | | |
| Basic | 0.099 | -0.039 | 0.024 | -0.012 | 0.007 | 0.047 | -0.034 |
| | (0.066) | (0.047) | (0.047) | (0.044) | (0.046) | (0.047) | (0.043) |
| Number of Observations | 69,197 | 140,619 | 139,596 | 130,723 | 136,084 | 131,095 | 184,445 |
| Number of Teachers | 1,547 | 3,062 | 3,013 | 2,918 | 3,188 | 3,155 | 4,677 |
| Dependent Var. Mean | 5.627 | 5.597 | 5.634 | 5.607 | 5.603 | 5.586 | 5.578 |
| Bandwidth | 0.115 | 0.172 | 0.186 | 0.150 | 0.185 | 0.167 | 0.317 |
| **Panel B: Lagged Indicator for Passing the School Year** | | | | | | | |
| Basic | 0.003 | -0.004 | -0.002 | -0.002 | -0.001 | 0.009 | 0.001 |
| | (0.010) | (0.007) | (0.006) | (0.007) | (0.006) | (0.006) | (0.013) |
| Number of Observations | 67,517 | 139,601 | 164,966 | 133,261 | 113,595 | 131,095 | 157,486 |
| Number of Teachers | 1,502 | 3,034 | 3,565 | 2,969 | 2,656 | 3,155 | 4,001 |
| Dependent Var. Mean | 0.960 | 0.961 | 0.962 | 0.962 | 0.962 | 0.965 | 0.963 |
| Bandwidth | 0.112 | 0.169 | 0.228 | 0.152 | 0.152 | 0.168 | 0.253 |
| **Panel C: Lagged Attendance Rate** | | | | | | | |
| Basic | 0.001 | 0.000 | 0.002 | -0.003 | 0.001 | 0.011** | -0.001 |
| | (0.005) | (0.005) | (0.005) | (0.007) | (0.005) | (0.005) | (0.008) |
| Number of Observations | 100,861 | 124,101 | 149,362 | 118,875 | 139,279 | 130,025 | 191,992 |
| Number of Teachers | 2,244 | 2,701 | 3,233 | 2,660 | 3,262 | 3,129 | 4,857 |
| Dependent Var. Mean | 0.931 | 0.930 | 0.928 | 0.921 | 0.920 | 0.918 | 0.913 |
| Bandwidth | 0.171 | 0.150 | 0.203 | 0.137 | 0.190 | 0.166 | 0.333 |

*Notes:* The sample is composed of all $4^{th}$, $8^{th}$ and $10^{th}$ grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between one and six years after. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable for whether the teacher responsible for that subject obtained a Basic score in his/her first evaluation, and the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year). I employ a fuzzy RD using a local linear regression. The dependent variables in Panel A and Panel B are the student's GPA in the previous year (1-10), and an indicator for whether the student passed the previous school year, respectively. In Panel C, the dependent variable measures the share of days the student attended school the year before. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the student's school are presented in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.7: Mean Baseline Characteristics for Subsamples Teaching a SIMCE Grade/Subject

| | Taught a SIMCE Subject/Grade (after first evaluation) in | | | | | |
|---|---|---|---|---|---|---|
| | Year 1 | Year 2 | Year 3 | Reev. Year | Year 5 | Year 6 |
| **Panel A: $4^{th}$, $8^{th}$ and $10^{th}$ Grade** | | | | | | |
| Male | 0.33 | 0.30 | 0.32 | 0.30 | 0.32 | 0.30 |
| Age | 45.33 | 44.30 | 45.00 | 43.92 | 44.44 | 43.54 |
| Years of experience | 17.94 | 16.67 | 17.61 | 16.25 | 16.93 | 15.84 |
| Has a degree | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.97 |
| Main job is teaching | 0.94 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 |
| More than one school | 0.10 | 0.09 | 0.10 | 0.09 | 0.11 | 0.11 |
| More than one municipality | 0.06 | 0.06 | 0.07 | 0.06 | 0.07 | 0.07 |
| Civil servant | 0.73 | 0.70 | 0.72 | 0.68 | 0.70 | 0.68 |
| Number of contract hours | 36.28 | 36.25 | 36.09 | 36.04 | 36.00 | 36.05 |
| Refused evaluation before | 0.02 | 0.03 | 0.02 | 0.03 | 0.02 | 0.03 |
| Overall evaluation score | 2.62 | 2.61 | 2.62 | 2.61 | 2.63 | 2.62 |
| Portfolio score | 2.23 | 2.23 | 2.23 | 2.24 | 2.24 | 2.26 |
| Lower primary school | 0.37 | 0.51 | 0.38 | 0.48 | 0.36 | 0.45 |
| Upper primary school | 0.48 | 0.39 | 0.47 | 0.40 | 0.48 | 0.34 |
| Secondary school | 0.15 | 0.10 | 0.14 | 0.11 | 0.15 | 0.20 |
| Urban | 0.62 | 0.60 | 0.63 | 0.61 | 0.63 | 0.64 |
| Enrollment | 491.37 | 460.26 | 490.50 | 464.89 | 495.07 | 521.94 |
| Teacher-student ratio | 27.81 | 27.34 | 27.93 | 27.52 | 27.88 | 28.17 |
| Average SES of students | 1.87 | 1.90 | 1.88 | 1.91 | 1.88 | 1.89 |
| School average SIMCE | 237.22 | 237.62 | 236.87 | 237.76 | 237.02 | 237.27 |
| School won previous SNED | 0.30 | 0.29 | 0.30 | 0.29 | 0.30 | 0.29 |
| Taught SIMCE grade/subject in year 0 | 0.11 | 0.33 | 0.11 | 0.37 | 0.15 | 0.25 |
| Students' average SIMCE in year 0 | 236.31 | 237.01 | 236.44 | 238.23 | 236.38 | 237.50 |
| Number of Teachers | 10,220 | 8,138 | 10,677 | 8,555 | 10,740 | 9,007 |
| **Panel B: $4^{th}$ Grade** | | | | | | |
| Male | 0.27 | 0.26 | 0.25 | 0.26 | 0.25 | 0.25 |
| Age | 44.55 | 44.37 | 44.30 | 44.27 | 44.28 | 43.88 |
| Years of experience | 17.03 | 16.62 | 16.81 | 16.65 | 16.70 | 16.24 |
| Has a degree | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| Main job is teaching | 0.90 | 0.91 | 0.92 | 0.91 | 0.91 | 0.91 |
| More than one school | 0.06 | 0.07 | 0.07 | 0.08 | 0.08 | 0.07 |
| More than one municipality | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.04 |
| Civil servant | 0.72 | 0.71 | 0.71 | 0.70 | 0.70 | 0.70 |
| Number of contract hours | 36.78 | 36.55 | 36.42 | 36.31 | 36.41 | 36.37 |
| Refused evaluation before | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| Overall evaluation score | 2.59 | 2.59 | 2.59 | 2.60 | 2.61 | 2.61 |
| Portfolio score | 2.20 | 2.20 | 2.20 | 2.22 | 2.22 | 2.22 |
| Lower primary school | 0.80 | 0.77 | 0.75 | 0.71 | 0.72 | 0.72 |
| Upper primary school | 0.20 | 0.23 | 0.25 | 0.29 | 0.28 | 0.28 |
| Secondary school | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Urban | 0.50 | 0.51 | 0.53 | 0.52 | 0.52 | 0.51 |
| Enrollment | 366.04 | 373.67 | 379.87 | 381.72 | 381.26 | 375.78 |
| Teacher-student ratio | 25.34 | 25.39 | 25.72 | 25.56 | 25.47 | 25.24 |
| Average SES of students | 1.85 | 1.85 | 1.87 | 1.87 | 1.86 | 1.86 |
| School average SIMCE | 237.31 | 236.76 | 236.98 | 237.41 | 237.03 | 237.52 |
| School won previous SNED | 0.28 | 0.28 | 0.28 | 0.29 | 0.29 | 0.30 |
| Taught SIMCE grade/subject in year 0 | 0.23 | 0.25 | 0.21 | 0.35 | 0.22 | 0.24 |
| Students' average SIMCE in year 0 | 236.08 | 236.79 | 236.65 | 239.60 | 237.22 | 238.42 |
| Number of Teachers | 4,496 | 4,668 | 4,835 | 4,583 | 4,496 | 4,301 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated by 2015. The sample under header *Year 1* is restricted to teachers who instruct a subject/grade that participates in SIMCE the year after their first evaluation. The analogous definition applies to the remaining columns. In Panel A, I consider teachers who instruct any of the SIMCE grades, whereas Panel B only considers $4^{\text{th}}$ grade teachers. All of the teacher school characteristics are measured at the year of the teachers' first evaluation, in the school in which they were evaluated, except for the school's average SIMCE score which is lagged by one year. *More than one School* and *More than one municipality* are dummies for whether the teacher worked in more than one school or municipality in the year of the first evaluation. *Civil Servant* is a dummy variable taking the value of 1 if the teacher has a civil servant position, and 0 if he/she is a contract teacher. *Main job is teaching* takes a value of 1 if the teacher's main position in the school involves teaching (as opposed to administrative duties). *Refused evaluation before* is a dummy variable taking the value of 1 if the teacher had refused to be evaluated before. *Average SES of students in school* is a 1-4 index measuring the average socioeconomic status of the school's students that participated in SIMCE in that year, and *School average SIMCE score* is the raw average score that these students got in that test. *School won previous SNED* is a dummy taking the value of 1 if the school won the previous edition of SNED. *Taught SIMCE grade/subject in year 0* is a dummy variable for whether the teacher taught a SIMCE grade and subject on the year of his/her first evaluation, and *Students' average SIMCE in year 0* is the students' average SIMCE score in this year, for the teachers who taught one of these grades and subjects.

Table A.8: Effect of Assignment to Remediation on Reevaluation Results – Quadratic Polynomial

| | Overall Score | > Basic | Portfolio | Peer | Supervisor | Self |
|---|---|---|---|---|---|---|
| **Panel A: Without Year FE and Controls** | | | | | | |
| Basic | 0.028* | 0.047** | 0.043*** | 0.034 | -0.050 | -0.003 |
| | (0.014) | (0.023) | (0.015) | (0.035) | (0.038) | (0.018) |
| Bandwidth | 0.353 | 0.403 | 0.394 | 0.387 | 0.398 | 0.391 |
| Number of Teachers | 27,900 | 30,537 | 30,081 | 29,791 | 30,239 | 29,959 |
| Dependent Var. Mean | 2.642 | 0.728 | 2.243 | 3.094 | 2.943 | 3.881 |
| **Panel B: Including Year FE and Controls** | | | | | | |
| Basic | 0.027** | 0.047** | 0.044*** | 0.026 | -0.049 | -0.004 |
| | (0.014) | (0.023) | (0.014) | (0.037) | (0.035) | (0.020) |
| Bandwidth | 0.387 | 0.411 | 0.415 | 0.339 | 0.444 | 0.329 |
| Number of Teachers | 28,507 | 29,461 | 29,686 | 26,032 | 30,734 | 25,412 |
| Dependent Var. Mean | 2.644 | 0.727 | 2.242 | 3.089 | 2.946 | 3.879 |

*Notes:* The sample is composed of all public primary and secondary school teachers in Chile who were evaluated for the first time in 2004-2010 and were reevaluated. The table presents the results of a fuzzy RD using a quadratic polynomial, where the dependent variable is specified in the column header, and the main independent variable is a dummy variable taking the value of 1 if the teacher obtained a Basic rating in his/her first evaluation. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. *Overall Score* is the final score that the teacher obtained in the second evaluation (the weighted average of the four instruments), and *> Basic* is a dummy variable taking the value of 1 if the teacher's score was above the Basic/Competent cutoff. *Portfolio* is the teachers' score in the portfolio, and *Peer*, *Supervisor* and *Self* are the scores in the peer assessment, supervisor assessment and self-assessment, respectively. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the teacher's municipality are presented in parentheses. The regressions in Panel B include year fixed effects and teacher and school controls measured at the year of the teachers' first evaluation (in the school where they taught during that year). The teacher and school characteristics are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES status of students, and whether the school won a teacher pay-for-performance tournament. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A.9: Effect of Assignment to Remediation on Students' Standardized Test Scores – Quadratic Polynomial

| | Number of Years After First Evaluation ($t_0$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Year $1^{st}$ Ev. ($t_0$) | $t_0$+1 | $t_0$+2 | $t_0$+3 | Year $2^{nd}$ Ev. ($t_0$+4) | $t_0$+5 | $t_0$+6 |
| **Panel A: Without Year FE and Controls** | | | | | | | |
| Basic | 0.097 | -0.021 | -0.027 | -0.060 | -0.127* | 0.137* | 0.042 |
| | (0.073) | (0.069) | (0.089) | (0.060) | (0.068) | (0.076) | (0.074) |
| Bandwidth | 0.218 | 0.259 | 0.224 | 0.271 | 0.284 | 0.262 | 0.405 |
| Number of Observations | 125,314 | 202,448 | 164,275 | 226,409 | 196,392 | 201,162 | 219,754 |
| Number of Teachers | 2,787 | 4,430 | 3,550 | 4,994 | 4,583 | 4,812 | 5,492 |
| **Panel B: Including Year FE and Controls** | | | | | | | |
| Basic | 0.044 | -0.003 | -0.041 | -0.050 | -0.100* | 0.113 | 0.020 |
| | (0.072) | (0.059) | (0.075) | (0.055) | (0.058) | (0.072) | (0.072) |
| Bandwidth | 0.166 | 0.228 | 0.204 | 0.245 | 0.279 | 0.232 | 0.248 |
| Number of Observations | 93,149 | 173,854 | 145,568 | 198,353 | 182,986 | 172,000 | 148,931 |
| Number of Teachers | 2,079 | 3,809 | 3,140 | 4,386 | 4,269 | 4,112 | 3,762 |

*Notes:* The sample is composed of all $4^{th}$, $8^{th}$ and $10^{th}$ grade public school students in Chile that participated in the SIMCE standardized test in 2004-2016 in math, language, and natural and social science, and had a teacher in the tested subject that was evaluated for the first time in 2004-2010 and was reevaluated after four years. I split the sample by the number of years that passed between the teacher's first evaluation and the students' standardized test, as indicated in the column headers, and only consider cases in which the students were tested on the same year as their teacher's evaluation ($t_0$), or between one and six years after. The running variable is the teacher's first evaluation score, centered around the Basic/Competent cutoff. The main independent variable is a dummy variable for whether the teacher responsible for that subject obtained a Basic score in his/her first evaluation, and the dependent variable is the student's standardized test score expressed as a z-score (i.e., standardized by grade, subject and year). I employ a fuzzy RD using a quadratic polynomial, without any controls or fixed effects in Panel A, and controlling for year fixed effects, subjectxgrade fixed effects, the baseline characteristics of teachers and their schools, and the SES and gender of students in Panel B. All regressions are conducted over the MSE-optimal bandwidth using a triangular kernel. Robust bias-corrected standard errors adjusted for clustering by the student's school are presented in parentheses. Teacher and school specific controls are measured at the year of the teachers' first evaluation (in the school where they taught during that year), and are age, gender, degree, years of experience, number of contract hours, type of contract, whether the teacher works in more than one school and/or municipality, whether teaching is his/her main job, fixed effects for the subject in which the teacher was evaluated, whether the teacher's main school is located in an urban area, whether he/she teaches in lower primary, upper primary or secondary school, whether the teacher refused to be evaluated before, the number of students in the school, the teacher-student ratio, the average SES of students, and whether the school won a teacher pay-for-performance tournament. Regressions in Panel B also include dummies for the student's gender and the average SES of students in his/her class. * significant at 10%; ** significant at 5%; *** significant at 1%

# Appendix B Attrition and Job Characteristics

Almost two thirds of teachers rated Basic or Competent were reevaluated by 2015. One of the main reasons for not taking the second assessment is retirement, or being close to the age of retirement and thus not having to participate in the evaluation. This is the case for almost 16% of the teachers in my sample rated as Basic or Competent in their first evaluation. There are four remaining sources of attrition that could potentially be affected by the results of the first teaching evaluation, but were not. These are leaving the school system, taking a job in a private school, taking an administrative position, and remaining in the public school system but postponing the reevaluation or refusing to take it.

Six percent of the teachers with a Basic or Competent rating were no longer in the school system after four years. This includes quitting or being fired, with the latter only applying to teachers with a temporary contract (one third of Basic or Competent teachers). It is uncommon for teachers with a permanent contract to voluntarily quit teaching, as only 3% leave in the four-year period after their first evaluation. It is thus unsurprising that this outcome is continuous at the cutoff between a Basic and a Competent rating. It is not as uncommon for teachers with a temporary contract to exit the school system by quitting or dismissal (the data does not distinguish). Almost 10% of the Basic/Competent teachers with a temporary contract were no longer working in a school four years after their first evaluation. However, since the emphasis of the Chilean evaluation system was to improve the competencies of Basic teachers instead of punishing them, it is plausible that quitting or firing decisions are not prompted by obtaining a Basic rating.[1]

The Chilean school system has three types of schools: public schools, subsidized private schools, and fee-paying private schools. The second source of attrition consists of teachers taking a job in a fee-paying private school or a subsidized private school. Since fee-paying private schools only account for 10% of all primary and secondary school teaching positions in 2005-2014, less than 1% of teachers rated Basic or Competent subsequently migrated to this sector. In addition to the small number of vacancies, teachers from public schools account for a very low share of the hires in this sector. For instance, public school teachers filled less than 5% of the vacancies for elementary school teachers in fee-paying private schools in 2005-2014. The situation is different in the case of private-subsidized schools, which employed almost 45% of all primary and secondary school teachers in this period (the same as the public school sector). Yet less than 5% of the teachers with a Basic or Competent rating were working in this sector four years after their first

[1] An important distinction arises in the case of teachers with an Unsatisfactory rating, who faced the threat of dismissal if they did not improve their performance in the following two evaluations. Turnover is higher among these teachers (14% of those with a temporary contract left the school system after four years). This is consistent with the findings of Dee and Wyckoff (2015), who show that teachers from Washington DC that faced the threat of dismissal after being evaluated were more likely to exit the school system.

evaluation. One possibility is that public school teachers do not apply for these jobs, as average salaries are higher in public schools than in private subsidized ones (Bravo-Urrutia et al., 2008). Furthermore, two thirds of the teachers in my sample had a permanent contract, and moving to the private sector would involve losing this job stability (Mizala and Romaguera, 2000). Teachers may still apply to private-subsidized schools due to the non-pecuniary aspects of the job, such as the type of students served by the school (Hanushek et al., 2004; Scafidi et al., 2007). Public schools house students from low-income families, whereas the subsidized private sector is more prevalent among middle-class families (Mizala and Romaguera, 2000). Nevertheless, either due to low demand or supply, a small share of hires in this sector come from public schools. For example, public school teachers filled less than 13% of the vacancies for elementary school teachers in private subsidized schools in 2005-2014. Another source of attrition involves teachers taking an administrative position in a public school, such as a school principal or the head of a technical pedagogical unit. However, only 3% of Basic/Competent teachers move on to an administrative role. Given how uncommon it is for public school teachers to be hired in private schools or to take on an administrative role, the lack of a statistically significant discontinuity in these outcomes at the cutoff between a Basic and Competent rating is not surprising.

The final source of attrition, affecting 6% of Basic or Competent teachers, consists of individuals who were teaching in a public school four years after their first assessment, but did not get evaluated. In a third of these cases, the teachers left the school system or took an administrative job the year after that. It is possible that these teachers postponed their evaluation in anticipation of not teaching in the public sector the year after. The remaining 4% of Basic and Competent teachers were not evaluated for unknown reasons. This accounts for a small share of teachers, as those who refuse to be evaluated are automatically granted an Unsatisfactory rating, as discussed in Section 2.

Focusing now on teachers who remained in the public school system and were reevaluated, it is important to understand why there are no differences at the cutoff in the likelihood of changing jobs, or in the type of school teachers work for. While it is rare for public school teachers to take a job in private schools, mobility across public schools is more common. Table 5 shows that 24% of teachers changed schools within the public sector between their first and second evaluation. Teachers may change jobs in search of better working conditions, such as lower commuting times or better-behaved students (Boyd et al., 2005; Hanushek et al., 2004; Scafidi et al., 2007), although salaries are slightly higher in schools that are isolated or serve low-income students (Mizala and Romaguera, 2000).[2] If schools use evaluation scores as an input in their hiring decisions, teachers

[2]The characteristic that most strongly correlates with whether a teacher changes schools is his/her degree of job stability. In particular, teachers with a temporary contract are 21 percentage points more likely to change jobs than teachers with a temporary contract. The likelihood of changing schools is also higher for teachers working in rural schools (9 percentage points) and teachers working in schools with students from a low socioeconomic status (7

who barely obtain a Competent rating may have a higher chance of getting hired to fill coveted positions. It is thus crucial to understand the process by which public schools hire teachers. Public school teachers are appointed by means of a public recruitment process organized by the municipal school authorities (Cabezas et al., 2017). Following the teacher statute, job vacancies are posted in national newspapers, and applications are reviewed by a recruitment committee formed by the head of the municipal school board, the school principal, and a randomly selected teacher. The teacher statute also states that applicants must be ranked according to their professional performance, seniority and training, although it does not specify how each of these aspects is to be measured or the weight it should receive. Qualitative evidence suggests that most recruitment committees shortlist candidates solely based on their curriculum, although teachers are sometimes required to submit references (Paredes et al., 2013). Candidates are then interviewed by the school principal, and in some cases by someone from the municipal school authorities. Although there is little evidence on the details of the hiring process, qualitative evidence from interviews to teachers suggest that evaluation results are not used as an input in the hiring process (Ortúzar et al., 2016). This is consistent with the fact that at the moment of reevaluation, there are no differences in the observable characteristics of the schools in which teachers with a Basic and Competent rating work for.

percentage points).