# Evaluating Teacher Evaluation: Evidence from Chile

Andreas de Barros*

**Structured Abstract for the SREE Spring 2019 Conference**

**Click here for the most current working paper version**

**JEL codes:** I21 Analysis of Education; I22 Educational Finance; I25 Education and Economic Development; I28 Government Policy; O12 Microeconomic Analyses of Economic Development.

**Keywords:** Chile; education; quasi-experimental methods; teacher evaluation.

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

# 1 Background/Context

While claims that performance evaluations lead to improved teacher productivity are controversial (Darling-Hammond et al. 2012), there is surprisingly little empirical research to inform related debates (see Taylor and Tyler 2012). Moreover, there are strong disagreements on whether evaluations should be solely based on teachers' ability to raise test scores, with frequent demands for more comprehensive, "formative" evaluation systems (see Grissom and Youngs 2016). Proponents of such formative approaches frequently refer to Chile's evaluation system as a best-practice example: For instance, a recent World Bank report concludes that teacher evaluation is "the essential backbone for a high-performing education system" and that, while "[p]utting in place a sound system of teacher evaluation is expensive and institutionally challenging", "Chile's comprehensive teacher evaluation system, Docentemas [sic], has shown that it can be done" (Bruns and Luque 2014, 215 et sq.).

# 2 Purpose/Objective/Research Question

This study provides causal evidence on the effects of formative teacher evaluations on teacher productivity (as measured by test score and non-test score outcomes). The paper moreover explores potential mechanisms by investigating effects on teaching behaviors and on teacher beliefs. To my best knowledge, this study thus provides the only second rigorous assessment of this relationship and the first analysis of a well-established evaluation system that operates at national scale.

# 3 Setting

This study leverages data for the universe of Chile's public-school students, in grade four, and their math and language teachers.

# 4 Population/Participants/Subjects

I restrict all analyses to teachers who were initially evaluated in elementary, and I drop those teachers who would have been too old to be eligible for re-evaluation two years later. I also drop the small share of approximately 1.8 percent of teachers with a performance rating that would have suggested an "unsatisfactory" rating. This renders a sample of 29,093 teachers who were initially evaluated in 2005-2013, of which 8,363 teachers were initially evaluated after the law was introduced, in 2011,

2012, or 2013. Of the initially evaluated teachers, 7,037, 6,883, and 5,721 teachers are observed in the three years after, teaching a total of 159,134, 153,724, and 129,827 4th-grade students, respectively.

# 5 Intervention/Program/Practice

Chile's national teacher performance evaluation system ("Docentemás") was introduced in 2003 as a formative, standards-based assessment system. In 2005, participation became mandatory for public school teachers. The evaluation includes four components with differing weights, as follows: A self-evaluation (10%), a third-party reference report (10%), a peer evaluator interview (20%), and a teacher performance portfolio (60%). The latter, in turn consists of a teacher's submission of a portfolio describing an eight-hour learning unit and of an announced video recording of a class. Sub-scores for each of these components are aggregated to a single, continuous performance score, which is then used to rate teachers along four performance levels (outstanding, competent, basic, and unsatisfactory).

# 6 Research Design

In 2011, Chile passed a new law, requiring teachers ranked in the "basic" (the bottom-second) performance category to be re-evaluated after two years (instead of four). My identification strategy exploits this variation across time, together with the discontinuity in the evaluation system's underlying scoring mechanism. I show that, although the newly introduced law is imperfectly observed, it sharply increased a "basic" teacher's likelihood to be newly evaluated after two years.

# 7 Data Collection and Analysis

The paper's analyses rest on data-sources with unusually comprehensive coverage of a national education system. For the years 2005 to 2015, I use teacher-classroom links to match data on the universe of elementary teachers in Chile's publicly funded schools, all teacher evaluations conducted in these years, administrative records for the universe of Chilean students, and results on standardized test scores for all Chilean $4^{\text{th}}$-graders in mathematics and reading.

The study's analytic findings rest on a difference-in-difference estimation strategy, which recovers the effect of teacher re-evaluations for those teachers who are assigned

to be re-evaluated and comply with their assignment (the "treatment-on-the-treated" or "ToT" effect). To confirm their robustness, these estimates are moreover compared to those using a novel difference-in-discontinuities ("dif-in-disc") estimator. For the purpose of this comparison, I extend a previously proposed difference-in-discontinuities ("dif-in-disc") estimator (Grembi et al. 2012) to the given case, in which the assignment mechanism is "fuzzy" (rather than "sharp"). Intuitively, this latter strategy thus takes the difference between a fuzzy regression-discontinuity (RD) estimate after 2011 and a respective estimate for the period before 2011.

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

# 8 Findings/Results

I cannot conclude that Chile's repeat performance evaluations lead to substantial gains in student achievement, one year after a teacher's re-evaluation, in both math and reading. However, for the year of the re-evaluation, I also do not find any negative effects on student learning. At the same time, I find that concerns about additional detrimental effects may be at least partly warranted: The paper's results suggest that re-evaluations led to decreases in teachers' level of caring, in the year prior to the re-evaluation. I do not find support for additional negative effects, as measured by teachers' beliefs in their students future educational attainment or their teaching practices. Analyses of effects on non-test score outcomes are currently ongoing (following Jackson (2018)); results will be presented at the Conference.

# 9 Conclusion

This study provides first evidence on the causal effects of formative teacher (re-)evaluations – under a comprehensive, standards-based teacher evaluation system that has recently been described as a role model for other countries (Bruns and Luque 2014). As discussed by Taut et al. (2011), Chilean policy-makers regularly re-consider whether Docentemás is worth its cost and whether the system should be expanded to private schools (Educación 2020 2013). Moreover, given the scale

and nature of the investigated program, even decision makers in other public sectors may look to the example of "Docentemás" as staff performance evaluation systems are (re-)considered. I am confident that such debates can be fostered by providing sound evidence on the effects of formative evaluations on teacher productivity (as measured by test score and non-test score outcomes), teacher beliefs, and teaching behaviors.

# References


Bruns, B. and J. Luque (2014, January). Great Teachers: How to Raise Student Learning in Latin America and the Caribbean. Technical Report 89514, The World Bank, Washington, D.C.

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

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

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

Grissom, J. A. and P. Youngs (Eds.) (2016). *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*. New York, NY: Teachers College Press.

Jackson, C. K. (2018, June). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*.

Taut, S., M. V. Santelices, C. Araya, and J. Manzi (2011, December). Perceived Effects and Uses of the National Teacher Evaluation System in Chilean Elementary Schools. *Studies in Educational Evaluation 37*(4), 218–229.

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