Economics of Education Review 58 (2017) 123–140

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico

Rafael de Hoyos[a], Vicente A. Garcia-Moreno[b], Harry Anthony Patrinos[c,*]

[a] *Lead Economist World Bank, 1818 H Street N.W., Washington DC 20433, USA*
[b] *Director General for Productivity Analysis, Ministry of Finance, Belen de las Flores Reacomodo Álvaro Obregón, Mexico City 86007, Mexico*
[c] *Education Manager, World Bank, 1818 H Street N.W., Washington DC 20433, USA*

ARTICLE INFO

Article history:
Received 22 November 2015
Revised 1 March 2017
Accepted 27 March 2017
Available online 7 April 2017

Keywords:
Accountability
Information
Education

ABSTRACT

The Mexican state of Colima implemented a low-stakes accountability intervention with diagnostic feedback among schools with the lowest test scores in the national assessment. A difference-in-difference and a regression discontinuity design are used to identify the effects of the intervention on learning outcomes. The two strategies consistently show that the intervention increased test scores by 0.12 standard deviations only a few months after the program was launched. The results indicate that full and wide dissemination of information detailing school quality is critically important.

1. Introduction

The increasing availability of standardized tests in developing countries opens the door to a wide range of policy interventions to improve student learning outcomes. Under certain conditions, the provision of information in itself can be conductive to an increase in learning outcomes. For instance, when test results are made public, they can eliminate heterogeneous perceptions among agents in the system (authorities, school directors, teachers, parents and students) regarding the quality of education services. Given sufficient levels of initial asymmetries on perceptions regarding the quality of education, the additional information brought about by test scores can be enough to generate a new equilibrium yielding a more transparent and accountable system providing higher quality education services. Information can affect positively school processes and outcomes because information promotes a dialogue and consultation among all actors. In the United States, some test-based accountability initiatives amounted to no more than information about school performance; yet, those experiences led to positive effects (Chiang, 2009; Loeb & Strunk, 2007; Reback, 2006).

The availability of test score information can also trigger actions to improve outcomes. This can happen through specific actions such as the development of school-specific improvement plans. If the standardized tests are universal in at least one school grade, then every school in the system could have access to a detailed diagnosis of the main challenges in the subject areas and grades assessed by the tests. If these detailed diagnoses and school improvement plans are followed by actions to address the problems identified, then the final outcome can be a system delivering higher quality services.

As shown in Bruns, Filmer, and Patrinos (2011), information for accountability can be a mechanism for improving school outcomes. There are three main accountability channels through which information could affect learning outcomes: increasing choice, participation and voice. There is evidence, mostly from high-income countries, showing that learning outcomes can improve as a result of more accountability. For instance, in the Netherlands, both average grades and the number of diplomas awarded increased after schools received a negative report card. For schools that received the lowest ranking, the one-year effects on final exam grades amounted to 10–30% of a standard deviation (Koning & van der Wiel, 2013). Evidence shows that schools in the United States respond to accountability pressures with improved test scores (Carnoy & Loeb, 2003; Hanushek & Raymond, 2005). Students in high-accountability states averaged significantly greater gains on 8th grade math tests than did students in states with little or no state measures to formally track student performance. One of the most prominent examples of accountability interventions is the state-level ranking system of No Child Left Behind (NCLB) in the United States. Rockoff and Turner (2010) found that the introduction of student tests and other measures to assign each school a grade was enough to increase student achievement in New York City. Rouse, Hannaway, Goldhaber, and Figlio (2013) show that

[*] Corresponding author.
E-mail address: hpatrinos@worldbank.org (H.A. Patrinos).

http://dx.doi.org/10.1016/j.econedurev.2017.03.007
0272-7757/

schools facing accountability pressures in Florida changed instructional practices and this, in turn, partly accounted for increases in test scores. Using administrative data for North Carolina, Ahn and Vigdor (2014) find evidence of a short term positive impact on school performance, and among low performing students in the medium term.

In most of the documented cases of education accountability interventions in high-income countries, school actors respond to this type of intervention when they are followed by rewards and/or sanctions. However, in low- and middle-income developing countries with a combination of relatively weak institutions, low managerial capacity of school directors, high levels of poverty and inequality, and the presence of powerful teacher unions, such high stakes accountability interventions might not be feasible nor desirable. Instead, a shared responsibility approach characterized by a supportive and collaborative environment rather than a punitive one might be a more effective accountability intervention. Although low-stakes accountability interventions like the one described here could work through implicit mechanisms, such as stigmatization or reputational damage, they could also have an impact through coordinated collaboration among school actors and pedagogical tools. Identifying the schools as low performing, meeting with them, developing a detailed diagnosis to identify their main challenges, and offering them advice to design a school-specific improvement plan, could be enough to improve service delivery.

The evidence on the effects of *low-stakes accountability interventions* within a supportive and collaborative environment in developing countries is mixed. The experimental evidence for India presented in Muralidharan and Sundararaman (2010) shows that a program that provided low-stakes diagnostic tests and feedback to teachers had no effect on student learning outcomes. In a randomized study in Punjab, Pakistan, Andrabi, Das, and Khwaja (2014) show that providing test scores to households and schools leads to increases in subsequent test scores by 0.11 standard deviations after one year of the intervention. Test score gains in public schools were in response to a low-stakes threat since there were no formal consequences attached to results. Public schools in Punjab face little competitive or regulatory pressure to perform, yet the information had a significant impact because there are non-monetary mechanisms such as social/community pressures on public school teachers that induce performance improvements if a school is revealed to have low test scores. In Latin America, the results presented in Mizala and Urquiola (2013) show that distributing information regarding schools' value added in Chile had no effects on enrollment, tuition levels or socioeconomic composition of students suggesting a limited effect of a low-stakes accountability intervention.

This paper evaluates the impact on test scores of a short-lived program suitable to measure the effect of a low-stakes accountability intervention within a supportive and collaborative environment. The program, PAE, short for *Programa de Atención Específica para la Mejora del Logro Educativo* was implemented in the Mexican state of Colima between January 2010 and mid-2011, with the objective of increasing learning outcomes among the worst performing public primary schools in the state. Although originally designed as a comprehensive schooling intervention with various components, PAE was cut short administratively and in the end only a subset of the components were implemented. Schools were informed of their test scores and that they were to be part of the program; once the program was launched and the list of PAE schools was publicly disseminated, participating schools were assigned a technical adviser who would visit the school and help diagnose the test score results and design a school improvement plan. Although similar programs have been evaluated, PAE focuses on absolute achievement, and for that reason, information on standardized achievement test results can generate more pressure on schools than information on value-added, as seen in the case of Chile (Mizala & Urquiola, 2013), perhaps because absolute achievement can be closer to what society and school administrators expect and value.

The present study follows two alternative strategies to identify the effects on PAE on test scores. The first is the difference-in-difference approach comparing the evolution in test scores between PAE and non-PAE schools over time. The second strategy exploits PAE's rigid eligibility rule (an exogenously determined cut-off point of the national standardized test) dividing schools into treatment and control groups and compares them through a regression discontinuity design. Both strategies consistently show that the intervention increased test scores by 0.12 standard deviations only a few months after program launch. PAE's positive effects are true for math and Spanish test scores and do not show statistical differences among boys and girls. However, closer inspection of the heterogeneity of effects shows that the program's impact was larger among students with relatively better initial conditions (i.e., students without an age-grade distortion).

The intuition behind these results is that public recognition of low performing schools, together with a detailed diagnosis of the school's main challenges and an invitation to network with other school directors, teachers and advisers to develop improvement strategies, is enough to improve the quality of education services. In fact, no additional inputs to the schools were funded or provided throughout the program. In this way, the current study examines how standardized tests were used to identify poor performing schools and how merely diagnosing school problems and developing a school improvement plan can lead to higher student performance.

The paper is organized as follows: Section 2 provides background information on the Mexican state of Colima, describes the PAE, and charts some trends in test scores. Section 3 details the methodology and identification strategies while Section 4 discusses the main results. Finally, a concluding Section 5 enumerates some policy recommendations.

2. Background and recent trends

Colima is a small state in the center-west region of Mexico, with 650,000 inhabitants, 34% of whom live in households with incomes below the official 2012 poverty line.[1] Since the decentralization of the education system that began in 1992, Colima has built an efficient school system adjusting the national educational programs to the state's specific characteristics and needs. Throughout the 1990s, Colima undertook innovative education policies such as the implementation and dissemination of one of the country's first standardized tests. By 2003, Colima outperformed all Mexican states in the OECD's PISA test and actually approached the OECD average. For example, Colima's math scores were on par with those for Greece and Serbia, and higher than those from Thailand, Brazil, Uruguay and Turkey.

At the end of school year 2005–2006, for the first time, the Federal Ministry of Education (SEP) applied the National Evaluation of Academic Achievement in School Centers (ENLACE), a universal standardized test. ENLACE, a low stakes test, gathered information annually on student performance in math, Spanish and a rotating subject for third, fourth, fifth and sixth graders in private and public primary schools in Mexico. By design, ENLACE had a national mean score of 500 and a standard deviation of 100 for every subject area and grade. In early October 2009, the results from ENLACE 2008–2009 were published, and Colima performed below the na-

[1] The poverty headcount ratio in Mexico was 45.4% in 2012.

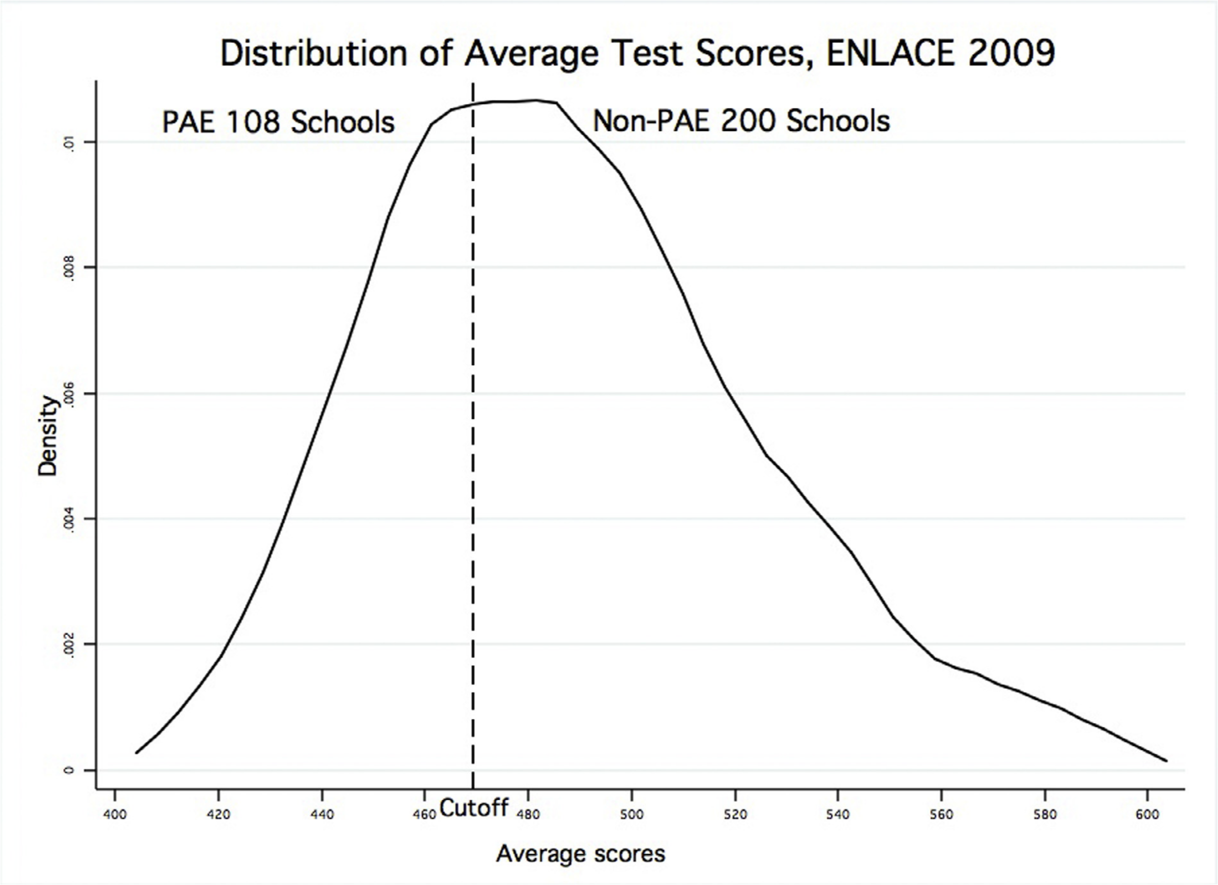


Fig. 1. PAE and Non-PAE population, Colima.

tional average. All the schools in Colima had access to this information, but there is no evidence of rankings or public dissemination of the results at this stage. A few weeks after the release of the disappointing ENLACE results, Colima's Ministry of Education (CMoE) began the design of PAE.

2.1. PAE

PAE was a program designed to improve learning outcomes among the lowest performing public primary schools in Colima, which also provide education services to students from the most marginalized schools. The program's operational rules excluded multi-grade schools with one or two teachers and community schools managed by the *Consejo Nacional de Fomento Educativo* (CONAFE).[2] Of the 477 primary schools in Colima in academic year 2008–2009, 40 were private, 39 were managed by CONAFE, 78 were one- and two-teacher schools, and 10, for a variety of reasons, did not have an ENLACE score. The group of PAE-eligible schools consisted of 310 public primary schools (see Fig. 1) with a total student population of 62,366 (95.2% of the total number of students in public primary schools in Colima during the 2008–2009 school year). Between October and November 2009, the CMoE used the 2008–2009 ENLACE score data to construct its ranking of schools. School scores were a simple average of the three subject areas tested: math, Spanish and science across grades 3, 4, 5 and 6. Schools in the 35th percentile or less of the distribution of the school average test scores were automatically designated as PAE schools (see Fig. A2 in Annex). As shown in Fig. 1, PAE included 108 of the 310 schools that belong to the potentially eligible population.[3] PAE schools were distributed across all ten municipalities of Colima and encompassed 1091 teachers and 10,550 students in 2009.

Fig. 2 illustrates the timeline followed by the design and implementation of Colima's PAE. Between November and December 2009, the CMoE assigned the schools that were going to participate in PAE following the criteria described above. In January 2010, the selected schools were officially notified. In February 2010, at a teachers' congress in Colima, the Governor launched the program and publicly disseminated the list of PAE schools. Although the assigned schools were presented as those with the lowest learning outcomes in the state, the Governor emphasized the co-responsibility of state authorities and the need to work closely with those schools to make improvements. Between the public announcement of the program and the first follow up ENLACE test in May 2010, PAE schools were assigned a *technical adviser*, part of the CMoE, who would visit the school three times a month to work with school directors and teachers on the diagnosis of the ENLACE test and the design of improvement strategies. In addition, the PAE technical adviser coached teachers on analyzing the ENLACE information to have a clearer understanding of how schools were assigned to PAE and the causes of poor performance within their schools.

Between January and March 2010, PAE' technical advisers, together with school directors and selected teachers, developed and applied a simple methodology to construct a detailed diagnosis identifying the academic weaknesses of their students based on

[2] For details on CONAFE see www.conafe.gob.mx.

[3] A total of 110 schools were originally selected to participate in PAE but two schools were dropped from the sample due to a mistake in their original classification as non-multi-grade schools which later on was changed to multi-grade.

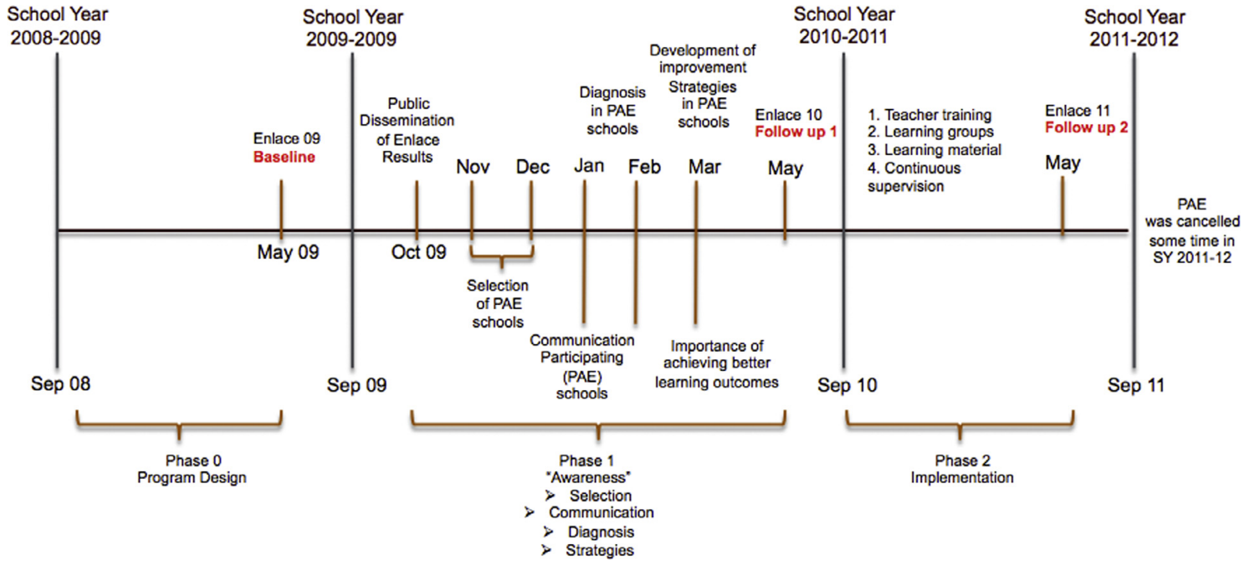


Fig. 2. The timeline of PAE.

ENLACE results.[4] The diagnoses were tailored to each school: the ENLACE test questions which more students answered wrong were collected by subject area, grade and classroom. Using personal identification numbers and a password, all teachers in Mexico had online access to a rich data set organizing the proportion of students in their class who answered an ENLACE question incorrectly. The website also indicated the area of knowledge and the relevant curriculum area for each ENLACE question, thereby providing teachers with concrete pedagogical direction to guide their efforts (see Fig. A1 showing an example of the type of information provided by ENLACE).

Between February and May 2010, PAE's technical advisers, school directors and teachers worked on a school improvement plan to address the problems identified during the diagnosis. The school improvement plan had to include clear medium- and long-term goals in terms of learning outcomes and a plausible strategy, involving teachers and parents, to reach them. The CMoE's technical advisers would visit PAE schools three times a month to follow up on the implementation of the school improvement plans. However, the "awareness" period of the program was too short to change any of the fundamental inputs of the learning production function and it is, hence, capturing the accountability and diagnosis effect of the program (see Fig. 2).

Implementation stage of PAE started in September 2010. It consisted of pedagogical interventions and the monitoring of progress. With diagnoses and school improvement plans in hand, state authorities, school directors and teachers collaboratively implemented the school-specific improvement strategies which, broadly speaking, included one or more of the following four interventions: (i) strengthening school-based management, (ii) training for school directors and supervisors, and (iii) reinforcing teachers' knowledge in the identified academic areas posing challenges. Due to reasons unrelated to the program's performance, PAE was canceled in mid 2011, less than one year after the pedagogical interventions were in place.[5] Given the partial implementation, at best, of the pedagogical interventions, the present study focus on the PAE's effects through accountability and availability of information mechanisms. A difference-in-difference and a regression discontinuity approach is used to answer the following two research questions:

1. Does PAE increase achievement test scores among students in participating schools?
2. Are PAE effects heterogeneous across students with different initial conditions?

2.2. Dataset and recent trends

This study uses and merges student test scores as measured by ENLACE with administrative school census data collected by federal and state education authorities (known as the *Formato 911*). Since 1998, this school census is collected at the beginning and end of each school year, and lists, among other entries, the number of teachers, students, classrooms, computers, the average years of schooling of teachers, and the geographic location of each school. With a unique school identifier (*Clave de Centro de Trabajo, CCT*), it is possible to merge this school census data with the results from ENLACE into a single data base.

ENLACE's methodology followed item response theory (IRT) allowing horizontal comparability of results (same grade over time), but not vertical comparisons (between grades). Since ENLACE is a census, in theory, it is possible to construct a panel of students with learning outcomes in different points in time. However, for this paper we do not attempt to construct and exploit the longitudinal dimension of ENLACE and instead rely on the analysis of repeated cross sections.[6] In addition to learning outcomes, ENLACE includes socioeconomic information for each school based on their geographical location.[7]

Fig. 3 shows mean math scores in PAE and non-PAE schools from 2006 to 2013. In general, schools in Colima improved by 42 points of ENLACE, or 0.42 standard deviations, throughout this period. As expected, the 108 PAE schools had lower learning out-

[4] The methodology relied on public information generated by the Federal Ministry of Education (SEP).

[5] SEP claimed that PAE was very close to the Federal Program "PEMLE" and to avoid duplicating efforts CMoE decided to discontinue PAE. Although the decision was confirmed in mid-2011, PAE did not receive enough budget to fully implement the pedagogical interventions during school year 2010–2011.

[6] More information on ENLACE is available from the test's website www.enlace.sep.gob.mx.

[7] The National Population Council (Consejo Nacional de Población, CONAPO) ranks all localities (an administrative and / or geographic entity often more disaggregated than a municipality) in Mexico according to a marginality index, a weighted average of literacy, access to basic public utilities, household infrastructure and average wages. Rankings range from very high marginalization, high marginalization, medium marginalization, low marginalization, and very low marginalization. For methodological details regarding Mexico's marginality index, see www.conapo.gob.mx.

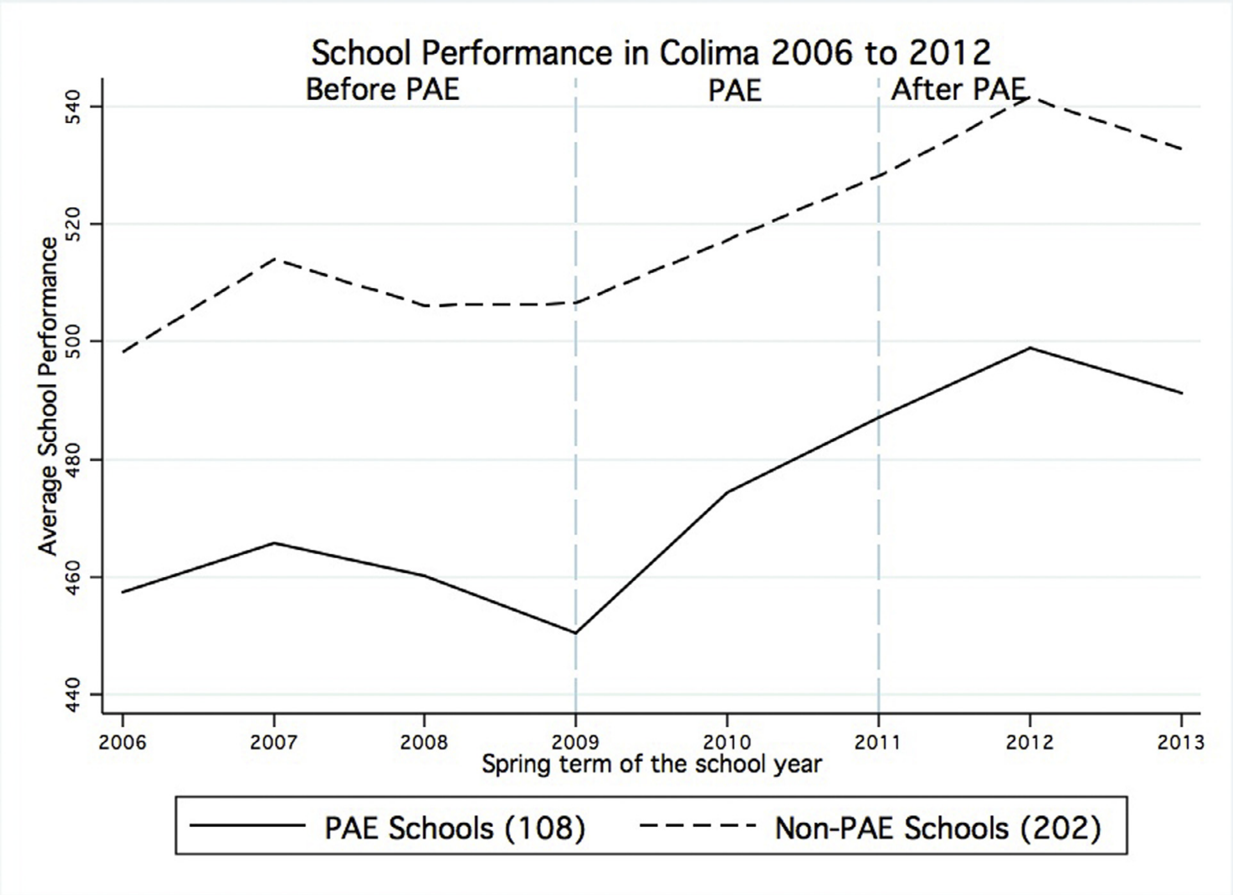


Fig. 3. Evolution of average score in PAE and non-PAE schools.

Table 1
Characteristics of public schools in Colima, 2009–2012.

Variable	School	2009	2010	2011	2012
Number of students	PAE	223.042	232.158	243.729	244.91
	Non-PAE	286.984	290.241	296.758	295.651
Student/teacher ratio	PAE	25.46791	26.38239	26.8327	26.89017
	Non-PAE	28.39154	28.35412	28.46782	28.40472
% of teachers with	PAE	0.4086632	0.4353684	0.4279682	0.4299641
Incentive program	Non-PAE	0.5917175	0.5751988	0.5549696	0.5519045
% of teachers with	PAE	0.7514346	0.7729504	0.7899677	0.7935849
B.A. or more	Non-PAE	0.7489002	0.7766308	0.8095469	0.8108789
Marginality index	PAE	1.191259	1.19247	1.188182	1.182734
	Non-PAE	0.4429964	0.4426403	0.4437509	0.4430926

Source: Authors' own computations with data from the school census, 2009, *Secretaría de Educación Pública.*

comes relative to the non-PAE schools. In 2009, the baseline year, the difference between PAE and non-PAE schools was 56 points; the gap was reduced to 42 point in 2010, one year after the implementation of PAE. Table 1 shows average characteristics of PAE and non-PAE schools during the period 2009–2012. PAE schools were smaller, with fewer teachers in the teacher incentive program (*Carrera Magisterial*) and significantly poorer than noon PAE schools.

3. Methodology

We follow two alternative methodologies to identify the effects of PAE on student-level test scores: a differences-in difference (DD) approach exploiting post-treatment differentiated performance on test scores between PAE and non-PAE schools and a regression discontinuity (RD) approach exploiting the exogenously defined threshold or cutoff point dividing PAE and non-PAE schools. The DD identifies the effects of the program assuming homogeneous performance in test scores between PAE and non-PAE schools in the absence of the program. In other words, DD assumes orthogonality between treatment assignment and pre-treatment trend, which could be a restrictive assumption. RD overcomes this limitation by exploiting differences in performance between PAE and non-PAE schools in the neighborhood of the exogenously determined cutoff point defining PAE schools. However, the limitation of RD is that the number of clusters (in our case schools) around the cutoff needed to identify a given effect under RD is relatively large, hence reducing the precision of the estimates. Therefore, getting similar effects under the two alternative strategies provides more robust evidence of the true impact of PAE on test scores. In addition, as a robustness check, we estimate the changes-in-changes estimator developed by Athey and Imbens (2006).

Formally, let us define $Y_{i,\,s,\,t}$ as the test score of the ith student in school s in year t and PAE_s as a dummy variable taking the value of one if the school is part of the program, zero otherwise. The DD is estimated via the following regression using OLS:

$$Y_{i,s,t} = \alpha_{0,t} + \alpha_1 PAE_s + A_t + \sum_{t=t^*}^{T} \lambda_t PAE_s * (A_t) + \sum_{k=1}^{K} \delta_{k,t} X_{s,t}^k + \epsilon_{i,s,t} \tag{1}$$

where A are year fixed effects, $X_{s,\,t}$ are a series of school-level controls and $\epsilon_{i,\,s,\,t}$ is a random component. The parameters of interest capturing the impact of PAE are the ones measuring the test score effects of the interaction between *PAE* and the year fixed effects, λ_t. If these parameters are statistically significant they would indicate that the post-treatment performance of PAE schools is different from non-PAE schools, controlling for everything else, suggesting that the program had an effect on student test scores.

There could be several reasons why an identification strategy such as the one described by Eq. (1) would yield biased estimators. First, the two groups, PAE and non-PAE schools are not equal ex-ante. Therefore, factors unrelated to the program can impact both groups differently causing a different post-PAE performance among treated and untreated schools and hence biasing the DD results. Second, the selection of schools on the basis of a one-year school performance ranking may misclassify schools due to a one-time performance aberration (one time shocks or mean reverting noise). As discussed by Chay, McEwan, and Urquiola (2005), this would produce biased estimators of the program's impact since PAE (low-performing) schools would tend to automatically revert to the overall mean, causing a post-treatment performance different from that of non-PAE schools.

According to Chay et al. (2005), a regression discontinuity design can defuse these two identification problems. The logic behind the RD approach is simple: the objective is to identify a group of schools that are part of PAE and similar enough to a group of schools that are not part of the program. A good place to identify such comparison groups is around the cutoff point distinguishing PAE from non-PAE schools as the threshold mimics a randomized selection to receive or not to receive treatment (Imbens & Lemieux, 2007; Imbens & Wooldridge, 2008). Formally, let us define $Y_{i,\,s,\,t}$ as a function of *PAE*, the average results of school *s* at the baseline $Y_{s,\,2009}$, the interaction between the former and the latter, a series of school-level controls $X_{s,\,t}$ and random component $\epsilon'_{i,s,t}$:

$$\begin{aligned} Y_{i,s,t} &= \beta_{0,t} + \beta_{1,t} PAE_s + \beta_{2,t} Y_{s,2009} + \beta_{3,t} PAE_s * (Y_{s,2009}) \\ &\quad + \sum_{k=4}^{K} \beta_{k,t} X_{s,t}^k + \epsilon'_{i,s,t} \end{aligned} \tag{2}$$

Notice that the dummy variable identifying schools belonging to PAE and their eligibility variable, ENLACE average results for 2009, are constant over time. By assumption $\epsilon'_{i,s,t}$ should be independently and identically distributed (iid) with a mean of zero and known variance. Eq. (2) can be modified to include higher order terms of the *forcing variable*, $Y_{s,\,2009}$, to control for non-linearities in the relationship between the eligibility criteria and subsequent test scores.[8] For a group of schools sufficiently close to the PAE-eligibility cutoff, such that samples are balanced both in observables and unobservables, the effects of PAE will be captured by $\hat{\beta}_{1,t}$ in Eq. (2). An important limitation to the regression discontinuity design, however, is that the results are valid only for observations around the cutoff point; the estimated impact is limited to a local average treatment effect which cannot be generalized to cover the entire population, thereby undermining the external validity of the estimation. A second limitation more relevant for the current study is that RD relies on having a large number of observations around the cutoff. As shown by Dragoset and Deke (2012) and Schochet (2009) for a constant statistical power and minimum detectable effect, the number clusters (schools) needed under RD is relatively large, hence, limiting the statistical power in the evaluation of PAE.

Table 2
Bandwidths around the cutoff.

Enlace points	Schools	Students	PAE	NonPAE
10.1	67	7460	35	32
20.2	129	14,126	67	62
40.4	223	25,071	98	125

Source: Authors' elaboration using RD command

Under both approaches, DD and RD, the unit of intervention is the school but the unit of analysis is the student (that is, schools, not students, are assigned to PAE). Therefore, the unobservables are composed of two terms $\epsilon_{i,s,t} = \eta_s + \nu_{i,s,t}$. These terms are, respectively, a school-specific component (η_s) and an individual-, school- and time-specific term $(\nu_{i,\,s,\,t})$. This structure of the error term implies clustering of students within schools allowing for intra-school correlation across students. Since PAE started in January of 2010 and the first follow up ENLACE test was a few months later, in May 2010, the DD estimator $\hat{\lambda}_{2010}$ and RD estimator $\hat{\beta}_{1,2010}$ capture the accountability effect of the program.

3.1. Determining the bandwidth of comparable schools

The optimal number of schools around the cutoff by which to evaluate the impact of the PAE program is determined by a trade-off between precision and internal validity. That is, a narrow bandwidth would select schools very close to the cutoff, hence more similar in observables and unobservables, but the statistical power might be compromised given the small number of observations. On the other hand, a wider bandwidth would increase the number of observations in the treatment and control groups but might not yield balanced samples in observables (and unobservables). We follow the method developed by Imbens and Kalyanaraman (2012) to determine the optimal bandwidth which yields 20.2 ENLACE points below and above the cutoff or 0.202 standard deviations around the threshold dividing PAE from non-PAE schools. This optimal bandwidth will be complemented with two alternative but rather arbitrary bandwidths: half of the optimal bandwidth (±10.1 points of ENLACE) and double the optimal bandwidth (±40.4 points of ENLACE).[9] Table 2 shows the number of schools above (non-PAE) and below (PAE) the cutoff as well as the number of students using each of the three different bandwidths.

A requirement for the RD approach to be valid is that the density of the *forcing variable* must be continuous around the cutoff and this is what is shown by Fig. A2 in the Annex. Granted sufficient observations around the cutoff, the RD approach may mimic a randomized experiment if the treatment and control groups are equal in expectation on all observed and unobserved dimensions. Table 3 shows school inputs in the school year 2008–2009 (the baseline) for schools within the optimal bandwidth. School inputs are statistically equal across treatment and control in all but one dimension. Differences in school size, number of teachers, the proportion of teacher that are part of a monetary incentives program (*Carrera Magisterial*), the proportion of teachers with a university degree or higher, as well as the dropout and failure rates between

[8] Higher order polynomials are typically used when estimating RD using all the available information as oppose to restricting it to those observations within the optimal bandwidth (see Imbens & Lemieux, 2007).

[9] The optimal bandwidth was computed using the regression discontinuity Stata program RD developed in Nichols (2014).

Table 3
School inputs 2009, schools within the optimal bandwidth.

	Non-PAE control	PAE-Treatment	Difference	S.E.
Number of students	201.9	181.7	20.17	(16.82)
Number of teachers	11.20	10.82	0.38	(0.73)
% of teachers with incentive program	0.51	0.45	0.06	(.055)
% of teachers with B.A. or more	0.72	0.75	−0.02	(0.04)
Student/teacher ratio	26.18	24.7	1.47	(0.89)
Marginality index	0.77	1.13	−0.36	(0.17)**

Source: Authors' own computations with data from the school census, 2009, *Secretaría de Educación Pública.*

PAE and non-PAE schools within the optimal bandwidth are statistically insignificant. Nevertheless, differences in the marginality index between PAE and non-PAE schools indicate that PAE schools tend to be poorer that non-PAE, even when comparing only those located ±20.2 points of ENLACE around the cutoff. This is not surprising given the very high correlation between test scores and socioeconomic status in Mexico and elsewhere (see Mizala, Romaguera, & Urquiola, 2007). By definition, PAE schools have lower average test scores than non-PAE schools, therefore, almost by construction PAE schools tend to be poorer than non-PAE schools. Our results control for all these school-level differences.

4. Results

This section shows the results of the DD and RD approaches described above focusing on the effects of PAE on math test scores. The effects on Spanish are listed in the Annex as a comparison. The results include estimations of DD and RD with and without school-level controls using all available data and restricting the estimation for those schools within the optimal bandwidth. The DD are also estimated using only schools around the cutoff to make them more comparable to the RD estimates and to make the results robust to differences in initial conditions. In all estimations standard errors are clustered at the school level.

4.1. DD approach

Table 4 presents the DD effects of the PAE program on math test scores using different specifications. The first of these specifications (column 1 in Table 4) estimates the DD effects without controls and using all available data of 310 PAE-eligible public primary schools in Colima. The results show an effect of almost 0.13 standard deviations (σ hereafter) in 2010, a few months after PAE was launched. According to this simple specification, the DD estimator increased over time although the estimated effects in 2011 and 2012 are not statistically different from that in 2010, as suggested by the learning outcome trends shown in Fig. 3. Notice that specification (1) with standard errors clustered at the school level is equivalent to specification without clusters but with school fixed effects. Specification (2) includes the following school-level controls: student - teacher ratio, proportion of teachers that are enrolled in the monetary incentives program *Carrera Magisterial*, proportion of teachers with a university degree or a post-graduate diploma, and the level of marginalization of the locality where the school is located (based on the marginality index). Including these controls, which are highly significant, reduces the estimated effects of PAE to close to 0.10 σ in 2010 but still significant at the 99% confidence level. The DD effects for 2011 and 2012 are positive and statistically significant but not different than the results in 2010.

Specifications (3), (4) and (5) are the same as specification (2) but restricting the observations to schools within the optimal bandwidth (OB), half of the OB, and twice the OB. Not surprisingly, under the OB restricted sample, the controls reduce their significance. However, the effects of PAE remain statistically significant with a point estimator close to 0.11 σ. Notice that under specification (3), the DD estimators across years 2010, 2011 and 2012 are very similar suggesting that there was a short term improvement and then test scores remained constant. Between 2009 and 2010, PAE schools managed to improve 0.11 σ faster vis-a-vis the changes in comparable non-PAE schools as a result of the intervention, capturing the effects of the program during the "awareness"

Table 4
Difference in difference PAE estimation, Math.

	(1)		(2)		(3)		(4)		(5)	
	All		All-controls		Optmal BW		Half the OB		Double OB	
PAE	−64.71***	(3.15)	−47.79***	(3.31)	−22.42***	(2.66)	−11.64***	(3.33)	−36.92***	(2.60)
2010	13.29***	(1.37)	13.88***	(1.38)	13.75***	(3.07)	16.18***	(4.54)	13.73***	(1.97)
2011	25.21***	(1.82)	26.34***	(1.82)	29.07***	(3.31)	30.40***	(5.36)	29.63***	(2.16)
2012	44.19***	(1.84)	45.52***	(1.85)	47.09***	(3.63)	51.34***	(5.05)	48.95***	(2.36)
PAE 2010	12.89***	(3.42)	9.63***	(3.51)	10.82**	(4.77)	10.93	(7.49)	10.31***	(3.61)
PAE 2011	18.47***	(3.30)	13.78***	(3.50)	12.85***	(4.81)	10.84	(7.28)	12.58***	(3.60)
PAE 2012	22.77***	(4.46)	17.62***	(4.55)	13.51**	(5.43)	7.88	(7.19)	15.92***	(4.77)
Student/teacher			1.97***	(0.39)	0.64*	(0.32)	0.49	(0.51)	0.52*	(0.30)
Incentive program			23.62***	(6.05)	8.70*	(4.92)	11.94*	(6.09)	5.41	(4.54)
Teachers BA^{n}			−5.47	(7.89)	−9.42	(7.03)	−9.07	(9.84)	−7.37	(6.09)
Low marginality^{n}			−12.42***	(3.87)	5.47*	(3.02)	5.16	(4.05)	4.13	(3.06)
Medium marginality^{n}			−16.02**	(6.33)	4.96	(7.68)	4.28	(8.76)	−3.34	(6.31)
High marginality			−15.52**	(7.12)	−1.88	(5.25)	−4.80	(5.00)	−4.40	(6.45)
Constant	523.28***	(2.69)	461.38***	(11.29)	473.47***	(9.60)	470.56***	(13.03)	487.68***	(8.84)
R2	0.063		0.075		0.038		0.037		0.050	
Obs	161,085		160,757		59,223		31,548		105,475	
Clusters	310		309		129		67		222	

School-level clustered standard errors in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.
n: the reference category is "very low" marginality level.

Table 5
Regresion discontinuity estimation results, math.

	(1)		(2)		(3)		(4)	
	All		Optimal BW		Half the OB		Double OB	
PAE	15.93^{*}	(8.91)	6.85	(8.59)	1.95	(13.61)	12.02^{*}	(7.24)
Forcing variable	0.98^{***}	(0.17)	0.44	(0.46)	0.23	(1.43)	0.99^{***}	(0.19)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	1.23	(0.93)	1.10	(0.76)	−0.06	(2.56)	0.38	(0.38)
$(PAE * (FV - cutoff))^2$	0.03	(0.02)						
Student/teacher	1.25^{***}	(0.35)	1.26^{**}	(0.51)	2.32^{*}	(1.26)	1.31^{***}	(0.42)
Incentive program	6.22	(5.24)	3.44	(6.81)	−6.47	(10.09)	−0.09	(5.79)
Teachers BA	−7.10	(7.18)	-24.84^{*}	(14.10)	-46.98^{**}	(20.03)	−13.01	(8.30)
Low marginality	2.43	(3.74)	2.96	(4.49)	12.01^{*}	(6.68)	4.76	(3.85)
Medium marginality	−4.06	(7.61)	7.40	(12.41)	22.12	(14.08)	0.74	(8.53)
High marginality	−1.54	(6.91)	0.86	(4.61)	8.55	(6.65)	−1.42	(6.35)
Constant	461.86^{***}	(11.77)	481.07^{***}	(18.81)	468.94^{***}	(29.26)	466.49^{***}	(14.18)
R2	0.095		0.012		0.014		0.030	
Obs	38,928		14,201		7518		25,433	
Clusters	307		127		67		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

period. Restricting the schools included in the analysis to half of the OB does not change the DD estimator; however, with only 67 clusters, the estimates lose precision and are no longer statistically significant. DD estimations using schools within double the size of the OB (222 clusters) yields very similar results as those with the OB. Albeit marginally smaller, the effects of PAE on math scores are corroborated by the program's effects on Spanish (see Table A1 in Annex). The effects on Spanish scores in 2010 range from close to 0.12 σ when estimating the DD without controls and using the full sample, to 0.09 σ when the controls are included and 0.08 σ when restricting the estimation to schools within the OB, always significant at the 95% confidence level.

4.2. RD approach

The first strategy followed within the RD approach is a graphical representation of the discontinuity using local linear or kernel regressions on both sides of the cutoff. Fig. 6 illustrates the relationship between average math performance at the school level in 2010 (vertical axis) and the *forcing variable*, the simple average ENLACE results at the school level in 2009 relative to the test score (horizontal axis). The PAE schools are to the left of the cutoff, which was used for their eligibility into the program, and the non-PAE schools are to the right. A mild discontinuity appears at the cutoff (469 points of the school average ENLACE score of 2009): there is a general pattern showing that schools slightly below the cutoff (the PAE schools) display greater test scores in 2010 than schools slightly above the cutoff (the non-PAE schools), although their scores in 2009 were very similar. Zooming into the discontinuity reveals that the difference in the regression at the cutoff is around 10 ENLACE points or 0.10 σ, very close to DD estimates. This graphic illustration suggests a positive effect on test scores in 2010 brought about by the program and is consistent with the results shown by the DD approach.

Analyzing the impact of the program after one year, when the schools may have implemented some of the pedagogical interventions, provides additional information on the impact of the PAE. The graphical representation of the RD in 2011, shows no discontinuity at the cutoff. Schools on the left of the threshold have an average achievement in 2011 similar to those schools on the right of the cutoff (see left panel of Fig. 6). Similar results showing no-discontinuity are found in 2012, two years after the implementation of the program.

The parametric estimation of RD, Eq. (2), includes the same specifications as the DD approach. In other words, the results presented in Table 5 include a specification using the full sample (more than 38,000 students in 310 schools at baseline) and including school-level controls while specifications 2, 3 and 4 restrict the observations to those schools within different bandwidths. All standard errors are clustered at the school level. According to the RD results of a specification using all available schools, PAE had an effect of 0.16 σ on math test scores in 2010 (with $\rho = 0.075$), a few months after the implementation of the program. These results are similar in magnitude to the ones obtained by the DD approach. However, the effect tends to disappear as we narrow the bandwidth of schools included in the regression. At double the OB (at ±0.4 σ around the cutoff) with 220 schools included, the effect of the program is 0.12 σ significant at the 10% level; at the OB (±0.2 σ around the cutoff) with 127 schools included the effect is no longer significant; and at half of the OB (±0.1 σ around the cutoff) with only 67 schools included, the effect basically disappears. These results suggest that PAE's effects could have been smaller among schools near the cutoff.

The parametric estimation of RD for years 2011 and 2012 are shown in Tables A4 and A5 in the Annex. Consistent with the graphic representation of the discontinuity, the parametric results for 2011 and 2012 show no effects of PAE on test scores. These results combined with the significant effects for 2010 suggest that the impact of PAE on test scores are driven by the *accountability* and diagnosis effect with no apparent short-run impact from the pedagogical interventions. It seems that the diagnosis and design of a school improvement plan based on the results of ENLACE was enough to improve test scores among PAE schools; however, no additional improvements were possible either because the pedagogical interventions were not able to address the constraints faced by schools or because the interventions were not well implemented, not implemented at all, or they simply needed more time to bear fruit.

While the DD and RD point estimates of 2010 are not substantially different from each other when all the schools in the sample are included or when those within double the OB are considered, the SE are significantly higher under RD. For instance, taking the full sample and including controls, the RD shows an impact of 0.16 σ with a standard error of 8.9, this last one being two a half times larger than the SE under DD which shows an effect of 0.10 σ and SE of 3.5. As we restrict further the bandwidth, SE increases and, particularly under RD, the point estimate reduces. The increase in

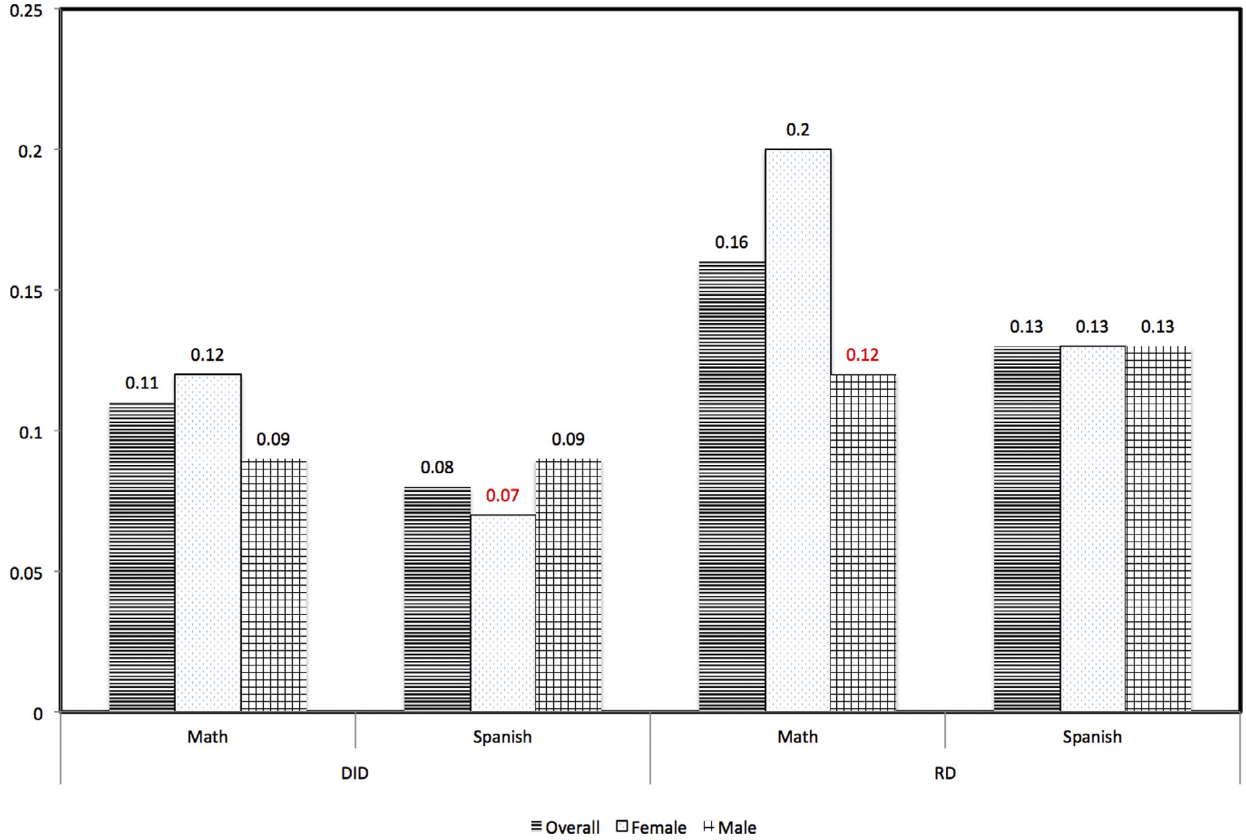


Fig. 4. PAE effects (in σ) by gender and subject, DID and RD, 2010. *Note:* Not significant coefficients in red, at 10% level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

SE or lost precision of the RD estimator can be explained by poor statistical power attributable to the relatively few schools around the cutoff. Notice that although there is a large number of students in our sample, the number of clusters or schools is relatively small, especially when we restrict the sample to schools located within the OB (127 schools). Schochet (2009) estimated that, for the same level of statistical precision, the number of clusters needed under RD are three to four times larger than the sample size required under a randomized controlled trial (RCT). According to Schochet (2009) "[t]he reduction in precision in the RD design arises due to the substantial correlation, by construction, between the treatment status and score variables that are included in the regression models; this correlation is not present under the random allocation design." Additional evidence in Dragoset and Deke (2012) shows that, under certain conditions, sample requirements are 9–17 times larger under RD that for RCTs, for a given statistical power. In our particular case, the number of clusters required under an RCT with an error of 0.05, power of 0.8, minimum detectable effect of 0.13 σ, an intra-class correlation of 0.05, R^2 of 0.05 and an average number of students per school of 250 is equal to 200 clusters, or 100 treatment and control schools, respectively. Taking the results of Schochet (2009), to have the same statistical precision under an RD we would need at least three times more clusters, or 600 schools within the OB, a figure considerably higher the 129 schools within the OB. Therefore, the lack of precision under RD when we restrict the sample to schools within the OB is explained by a lack of statistical power due to the low number of schools around the cutoff.

4.3. Heterogeneous effects

The average positive and significant effect of PAE could hide important heterogeneous impacts within schools. The program could have a differentiated effect among students with different initial conditions. For instance, between 2009 and 2010, the student-level standard deviation of test scores in Colima increased by 4 ENLACE points while the increase was equal to 10 points among PAE schools. This suggests that PAE could have increased the within-school dispersion in test scores due to a heterogeneous effects among students in different points in the distribution of skills.

To explore possible heterogeneous effects of the program, separate specifications are estimated for boys and girls and for students identified as having an age-grade distortion and those who do not. All the result presented in this section concentrate in the period 2009–2010 when the positive effect of PAE is observed. Fig. 4 shows PAE's test score effects in math and Spanish, differentiated by gender using the two alternative methodologies, DID and RD. The results show that although the program had some heterogeneous impacts across boys and girls with boys experiencing larger improvements in math and girls in Spanish, we cannot reject the null hypothesis of equality of coefficients between gender-specific equations.

Students with an age-grade distortion in Mexico are largely the outcome of late enrollment in the education system. As shown by Manacorda (2012) age-grade distortions have significant and long-lasting negative effects on test scores. In 2009, more than 17% of the students in Colima's public primary schools had an age-grade distortion and close to 40% of them fell in the "insufficient" level in ENLACE compared to 20% among students without an age-grade

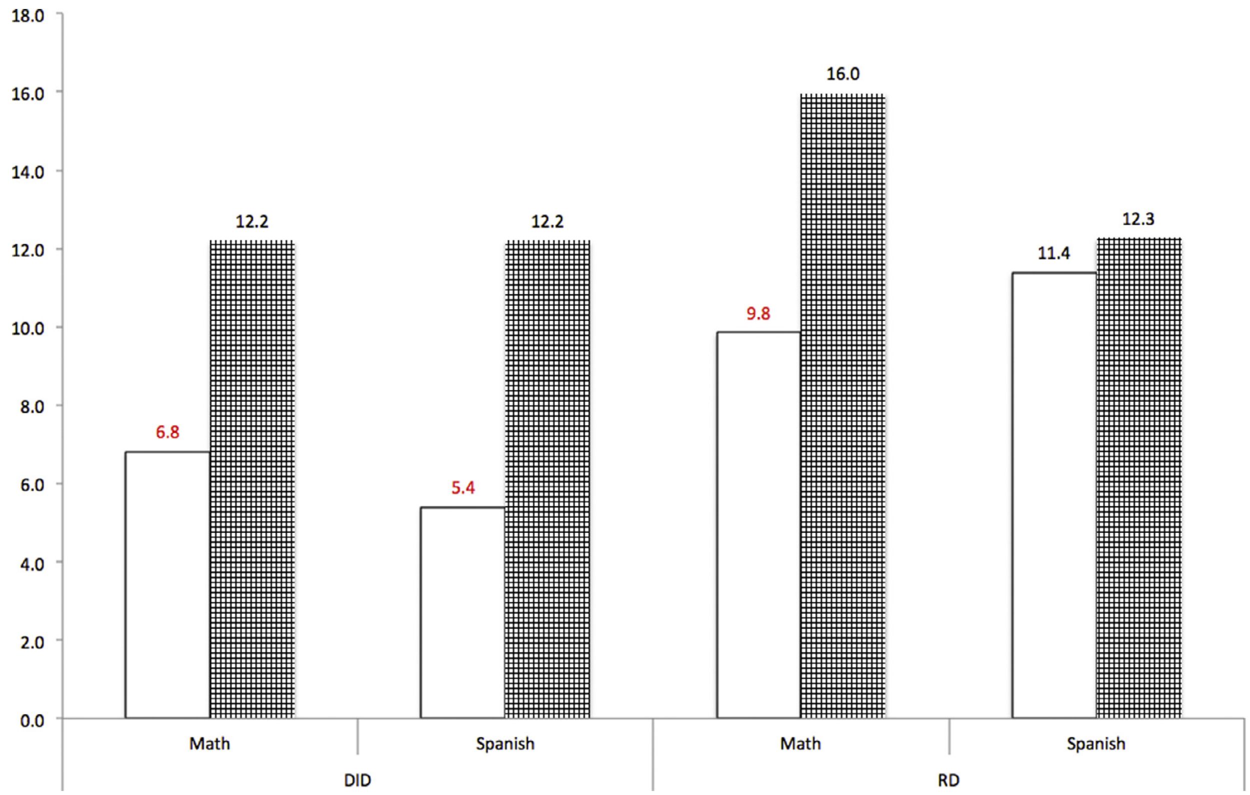


Fig. 5. PAE effects (in σ) by age-grade distortion and subject, DID and RD, 2010. *Note:* Not significant coefficients in red, at 10% level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distortion. Fig. 5 shows PAE's test score effects in math and Spanish, for students with an age-grade distortion and those who are on the right grade based on their age, using the two alternative methodologies, DID and RD. The results show that PAE had a positive effect, though not statistically significant, on math and Spanish test scores among students with an age-grade distortion while the effect was positive and significant among students without a distortion. This has two important implications: (1) the improvements on average test scores brought about by PAE were not the outcome of an increase in results among the better-off students while reducing achievement test scores among the relatively worse-off, and (2) students with relatively better starting conditions seem to benefit more from the changes introduced by a low-stakes accountability and diagnosis interventions such PAE. Arguably, improving scores among students with more challenging starting conditions requires more comprehensive and long-term interventions.

4.4. Robustness

The results presented so far suggest that PAE had a positive and significant effect on learning outcomes in the very short term. There are at least three threats to the internal validity of these results: (1) serially correlated outcomes leading to the underestimation of standard errors under DD; (2) a mean-reversion process artificially increasing the point estimator under the DD approach; and (3) differences in the distribution of unobservables across treatment and control groups, making the DD estimates inconsistent. This section address these three potential threats to the validity of our results.

Bertrand, Duflo, and Mullainathan (2004) show that, by ignoring the fact that DD focuses on serially correlated outcomes, its estimation can severely underestimate standard errors. They show that a simple correction consisting of collapsing the time series information into a "pre"- and "post"-intervention period would be enough to account for this time series inconsistency in standard errors. Tables A2 and A3 in the Annex shows that when the DD is performed using a sample with data collapsed into two periods, before and after PAE, the results are practically unchanged. The effects on math (Spanish) test scores range from 0.18 σ (0.14 σ) with the full sample and no controls to 0.14 σ (0.09 σ) when the controls are added and 0.13 σ (0.08 σ) when the sample is restricted to schools within the OB, in all cases significant at the 95% confidence level. As mentioned in Section 3 one of the most important limitations of the DD approach as a strategy to evaluate the impact of PAE is that program effects can be confounded with mean-reversion noise caused by a treatment selection based on ranking of schools at the baseline. Although the RD is a robust strategy to address this concern, we also perform a simulation showing that mean-reversion is not driving our results. If mean-reversion is at work, then reproducing PAE's eligibility criteria but using the ENLACE results of 2008 (as opposed to 2009) to rank and select schools should result in the simulated PAE schools advancing faster than the simulated non-PAE schools between 2008 and 2009. Fig. A3 in the Annex shows the average evolution of test scores between PAE and non-PAE schools using the alternative selection criteria based on the school ranking of 2008 (top figure) and compares it with the evolution under the actual selection of PAE and non-PAE schools. The simulation and actual trends on average test scores in Fig. A3 are based on schools within the OB. Using the ranking of 2008 to simulate the assignment of schools into PAE and non-PAE produces no differentiated evolution in test scores after the assignment, suggesting that the post-2009 PAE versus non-PAE differences in trends observed after the program was launched can be attributable to the intervention with little or no evidence of mean-reversion effects.

Differences in the pre-program distribution of unobservables across treatment and control groups would result in inconsistent

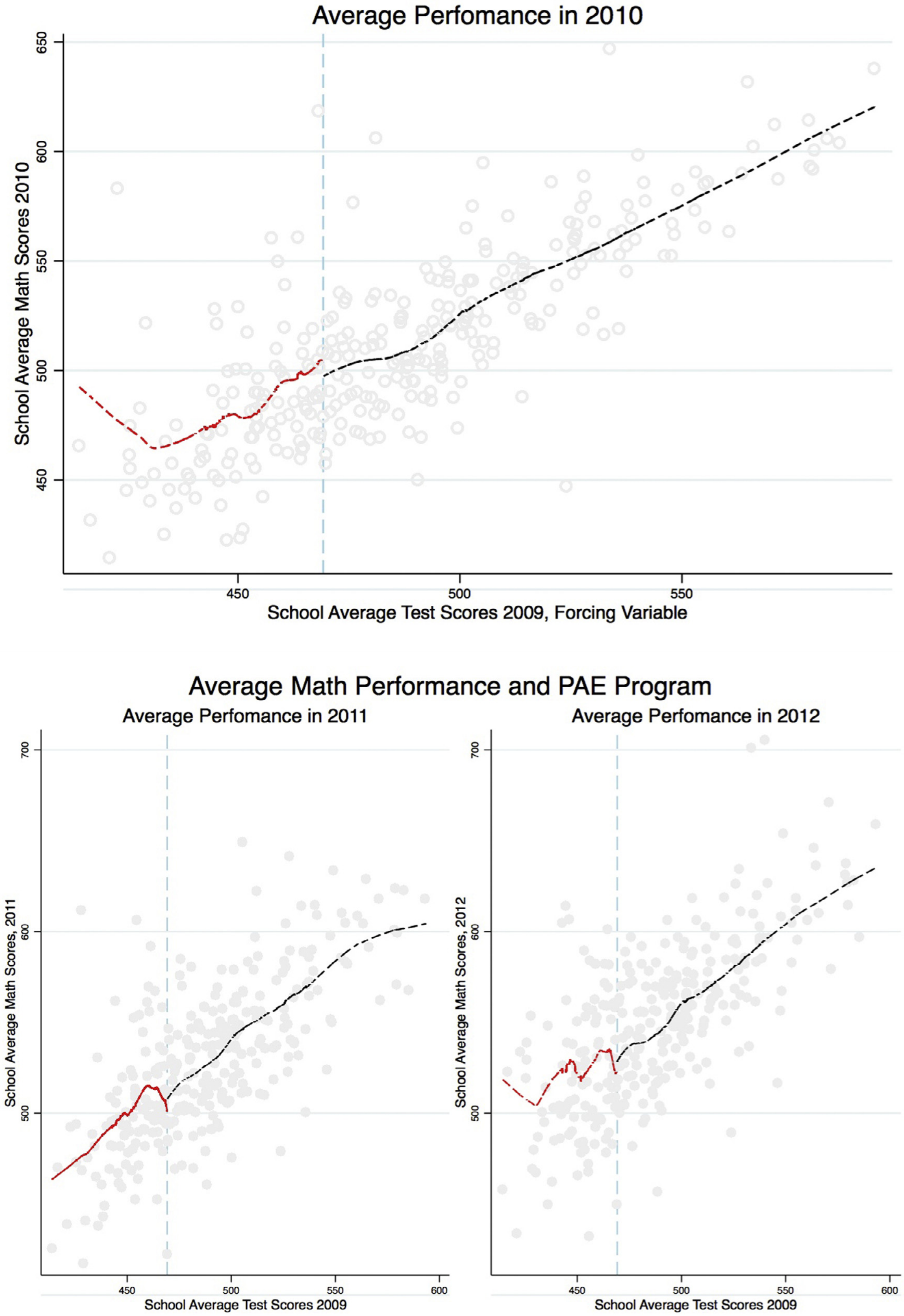


Fig. 6. Regression discontinuity, PAE vs Non-PAE, 2010, 2011 and 2012.

Table 6
Treatment effects under the changes in changes.

	Math				Spanish			
	Full sample		Optimal BW		Full sample		Optimal BW	
1	17.1	(3.44)	13.8	(3.70)	16.6	(2.95)	13.0	(3.16)
2	13.1	(2.89)	9.1	(3.25)	16.3	(1.78)	15.1	(2.68)
3	16.0	(2.71)	13.7	(3.55)	16.4	(1.99)	15.5	(2.94)
4	17.7	(2.61)	12.4	(3.83)	16.1	(2.36)	14.1	(2.47)
5	18.0	(2.60)	14.6	(3.63)	20.2	(2.41)	11.7	(3.62)
6	22.6	(2.90)	18.7	(3.78)	16.5	(2.75)	9.5	(3.85)
7	22.0	(3.14)	21.0	(4.16)	16.4	(3.04)	6.3	(3.92)
8	21.3	(3.14)	18.5	(3.93)	14.7	(3.28)	7.9	(3.43)
9	22.6	(3.05)	17.3	(3.98)	13.2	(3.09)	5.6	(5.07)
Average	18.9		15.5		16.3		11.0	

Standard errors in parentheses.
Standard errors estimated by bootstrapping results 1000 times.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

DD estimates. To address this potential limitation we complement the DD with the *changes-in-changes* (CC) estimator developed by Athey and Imbens (2006).[10] The CC estimator is a generalization of DD but imposing less assumptions. Under CC, the pre-PAE distribution of unobservables are allowed to differ between treatment and control groups. These pre-program distribution of unobservables are used to estimate the post-program treatment effect, giving a more robust estimate of the true treatment effect of PAE. A second advantage of CC over DD is that the former allows for heterogeneous treatment effects across the baseline distribution of outcome variables. In our case, CC can estimate the effects of PAE at different points of ENLACE test scores of 2009 (the baseline).

The results of the CC under different specifications, for math and Spanish are presented in Table 6.[11] Using the full sample of 310 schools without covariates, the CC effects show that PAE had a positive and statistically significant effect on math (0.19 σ on average) and Spanish (0.16 σ) test scores. These effects are larger than the ones estimated by DD and RD, perhaps indicating that the pre-program distribution of unobservables had a negative effect on test scores of PAE schools relative to non-PAE schools. Once these pre-program differences in unobservables are accounted for, the effects of PAE are indeed larger than the effects estimated under DD and RD. The difference in unobservables' effects among treatment and control groups could be explained by PAE's students relatively poorer background versus students in non-PAE schools. When the sample is restricted to schools within the OB, the average CC estimates are smaller but still statistically significant, both for math and Spanish. The heterogeneity of effects across the deciles of the baseline ENLACE score corroborates that students with relatively better initial math test scores at baseline benefited more from PAE. However, this was not the case in Spanish with students in the lower or middle part of the baseline test score distribution benefiting relatively more than their better-off peers.

PAE's information dissemination strategy could have created incentives within schools that can explain a positive impact. First, the pressure put on PAE teachers by being declared low-performing schools may have created incentives to practice some type of "strategic behavior" by school directors or teachers such as cheating or teaching to the test (see Figlio & Getzler, 2002). ENLACE uses two algorithms to detect cheating and results are invalidated when it happens.[12] There is no evidence of test scores invalidation by the Ministry of Education of Colima to any PAE school during 2010, 2011 and 2012. In addition, the percentage of students who did not take ENLACE in 2010 and 2011 was equal to share in non-PAE schools, which suggests that PAE schools did not try to manipulate test scores by choosing the students who took the test. Second, student mobility across schools might have affected the test scores after the PAE schools were identified. But such mobility in Mexico is very difficult without a strong reason such as the geographic reallocation of the student's family. In addition, it would have been more likely that the best students would have moved out of PAE schools, which, if anything, would suggest the impact of the program could have been stronger than the estimated presented in this study. Third, school directors could have changed their managerial practices in PAE schools. There is some evidence, from self-reported surveys, that directors in PAE schools improve monitoring of teacher's attendance and punctuality, visited classrooms more often, and had meetings to discuss learning outcomes between 2009 and 2010; all these managerial changes were prompted, in part, by the regular visits of PAE's *technical adviser*. Finally, principals and teachers may have focused on teaching to the test. Curriculum and ENLACE are linked by design and ENLACE was used to show the weakness areas in the classrooms of the low-performing schools. With the existing information, we cannot rule out the possibility that teachers used this information to teach the subjects that were more closely connected to the test (math and Spanish). But just as probable as using the information to improve learning outcomes, as argued here. There is little incentive to teach to the test since the Mexican achievement tests are not high stakes. In the case of the Colima intervention, the targeted schools were low performing, significantly below the state and national averages. Incentives to improve learning outcomes in this low-stakes environment probably resulted in increased attention to mastering the material, which is a positive thing when learning outcomes are so low.

5. Conclusions

In 2009, the state of Colima identified 108 public primary schools that had obtained the lowest learning outcomes as measured by the national standardized student assessment, ENLACE. In early February 2010, the state governor announced the "performance status" of selected schools: schools which performed below an arbitrary cut-off were automatically enrolled in a mandatory school improvement program known as PAE. The program, however, was discontinued during the 2011–2012 school year.

Following two alternative strategies to identify the effects of PAE on learning outcomes, a difference-in-difference and a regression discontinuity design, the paper shows that PAE increased test scores by 0.12 standard deviations only a few months after program launch. The size of the effect aligns with other studies evaluating the impact of low-stakes accountability interventions on test scores. Although this effect remains two years after program implementation, our results show no additional impact attributable to the interventions intended to change major inputs in the learning production function as oppose to a marginal change in effort. The effects are homogeneous across boys and girls; however, test scores among students with disadvantaged initial conditions, proxied by age-grade distortion, improved only marginally (statistically not different from zero) as a result of the intervention. Our results are not driven by serial correlation or mean reversion effects and are robust to the less restrictive identification assumptions under the changes-in-changes estimator.

[10] The authors are indebted with an anonymous referee who suggested the use of the CC estimator as a strategy to provide more robust evidence of the true impact of PAE.

[11] The CC effects were estimated using the Stata command "cic" developed by Blaise Melly at the University of Bern.

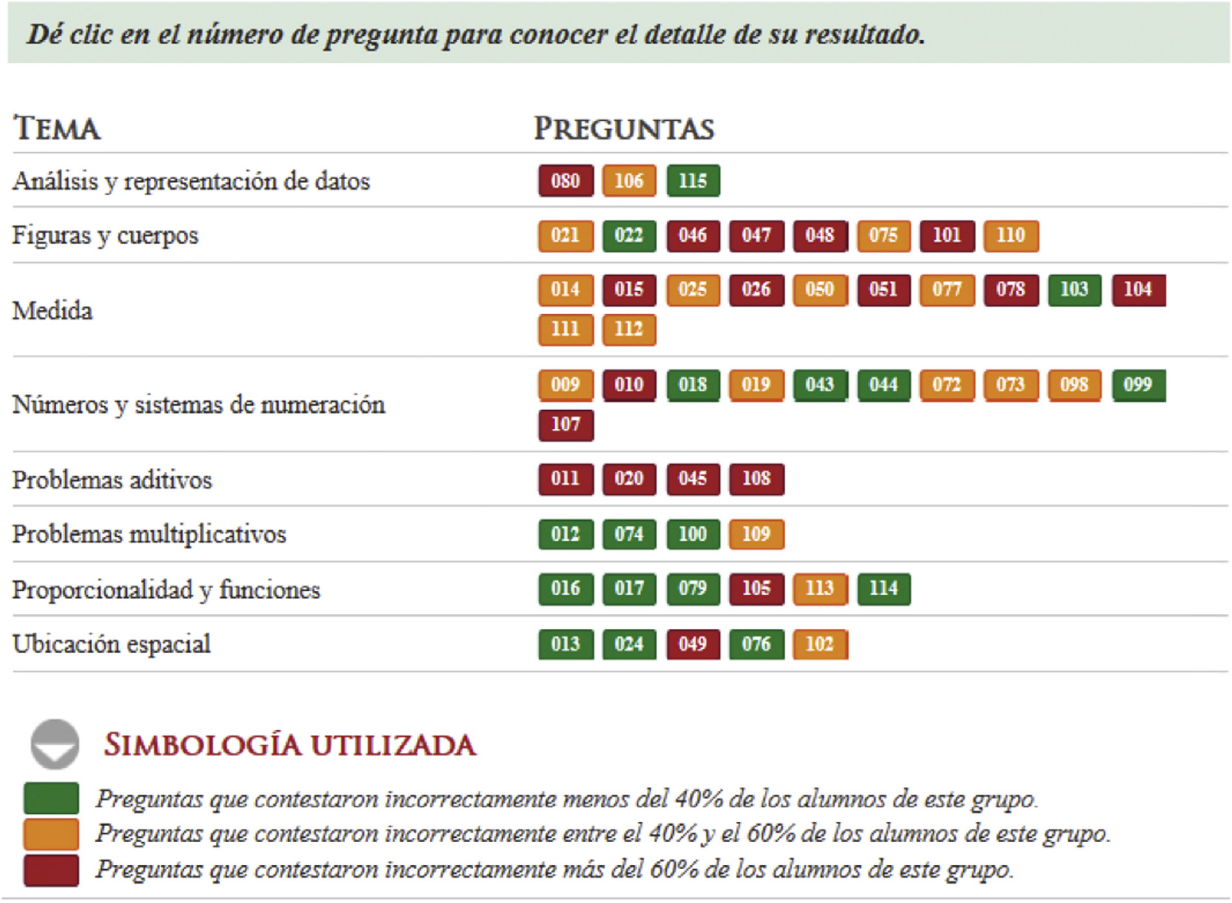
[12] Algorithms have also been used in the US to detect cheating, see Jacob and Levitt (2003) for an example using data from Chicago public schools.

The fact that the PAE program was halted after only 18 months of implementation suggests that the main intervention of the program was circumscribed to the public announcement made by state authorities, followed by detailed information provided to the schools about the test scores of their students, the activities connected to the design of a school improvement plan, and close support provided by the program's technical advisers. Activities during the period of preparation of the school improvement plan included the notification to schools that they were low-performing, a diagnosis based on test score results, identification of weaknesses within subject areas evaluated, a discussion between the school director and teachers on how to address the challenges, and the setting of clear goals regarding learning outcomes. In other words, it was the information that was publicly announced apprising directors in PAE schools of their relatively poor performance. While this information was public already, the announcement by state authorities triggered an accountability effect. The diagnostic feedback that came about through the design of the school improvement plan gave the schools the tools and knowledge they needed to take action and set goals themselves. Therefore, it is plausible that the public announcement itself allowed school to make small but significant learning gains.

The results suggests that when students, teachers and parents in a school know that their scores are low, this could trigger a process of self-evaluation and analysis, and the process itself may lead to an improvement in learning outcomes. Although there was no "shaming" for PAE schools in Colima, there may be an intrinsic motivational impact connected to the ranking of a school relative to others compounded by the compensatory nature of the program and the co-responsibility of state authorities in the challenge of improving learning outcomes. According to this analysis, it is not the inputs made available by PAE that led to improvements. Rather, it was the signaling value of the program which resulted in rising test scores. Moreover, unlike the high-stakes accountability interventions sometimes leading school closures in the United States, or the sacking of school directors in England, or the lead with your feet school choice in the Netherlands, the policy (and the *de facto* events) in Colima bore no punitive actions against schools or school directors.

While the PAE program in Colima was surprisingly and frustratingly short-lived, its premature termination serves to highlight a largely unrecognized phenomenon in education: acknowledgment is, in some ways, virtually tantamount to improvement. After all, if you really understand the problem, effective solutions come much easier. If you do not understand the problem, no amount of "problem-solving" can be expected to work. One may still legitimately wonder why schools did not improve before the PAE program given that the same information was already disclosed publicly. Perhaps the information was not well understood or disseminated, or beleaguered school leaders in poorly performing schools could not, without the right logistical support and networking, begin to proactively use the results from the standardized test to trigger a discussion and design a school improvement plan. These are all areas of future research. It remains refreshing, however, that the use of information from standardized tests, without punitive measures but within a supportive and collaborative environment, appears to be sufficient for improving learning.

Annex



Dé clic en el número de pregunta para conocer el detalle de su resultado.

TEMA	PREGUNTAS
Análisis y representación de datos	🟥 080 🟧 106 🟩 115
Figuras y cuerpos	🟧 021 🟩 022 🟥 046 🟥 047 🟥 048 🟧 075 🟥 101 🟧 110
Medida	🟧 014 🟥 015 🟧 025 🟥 026 🟧 050 🟥 051 🟧 077 🟥 078 🟩 103 🟥 104 🟧 111 🟧 112
Números y sistemas de numeración	🟧 009 🟥 010 🟩 018 🟧 019 🟩 043 🟩 044 🟧 072 🟧 073 🟧 098 🟩 099 🟥 107
Problemas aditivos	🟥 011 🟥 020 🟥 045 🟥 108
Problemas multiplicativos	🟩 012 🟩 074 🟩 100 🟧 109
Proporcionalidad y funciones	🟩 016 🟩 017 🟩 079 🟥 105 🟧 113 🟩 114
Ubicación espacial	🟩 013 🟩 024 🟥 049 🟩 076 🟧 102

SIMBOLOGÍA UTILIZADA

🟩 *Preguntas que contestaron incorrectamente menos del 40% de los alumnos de este grupo.*
🟧 *Preguntas que contestaron incorrectamente entre el 40% y el 60% de los alumnos de este grupo.*
🟥 *Preguntas que contestaron incorrectamente más del 60% de los alumnos de este grupo.*

Fig. A1. Example of a report card using ENLACE, math 3rd grade.

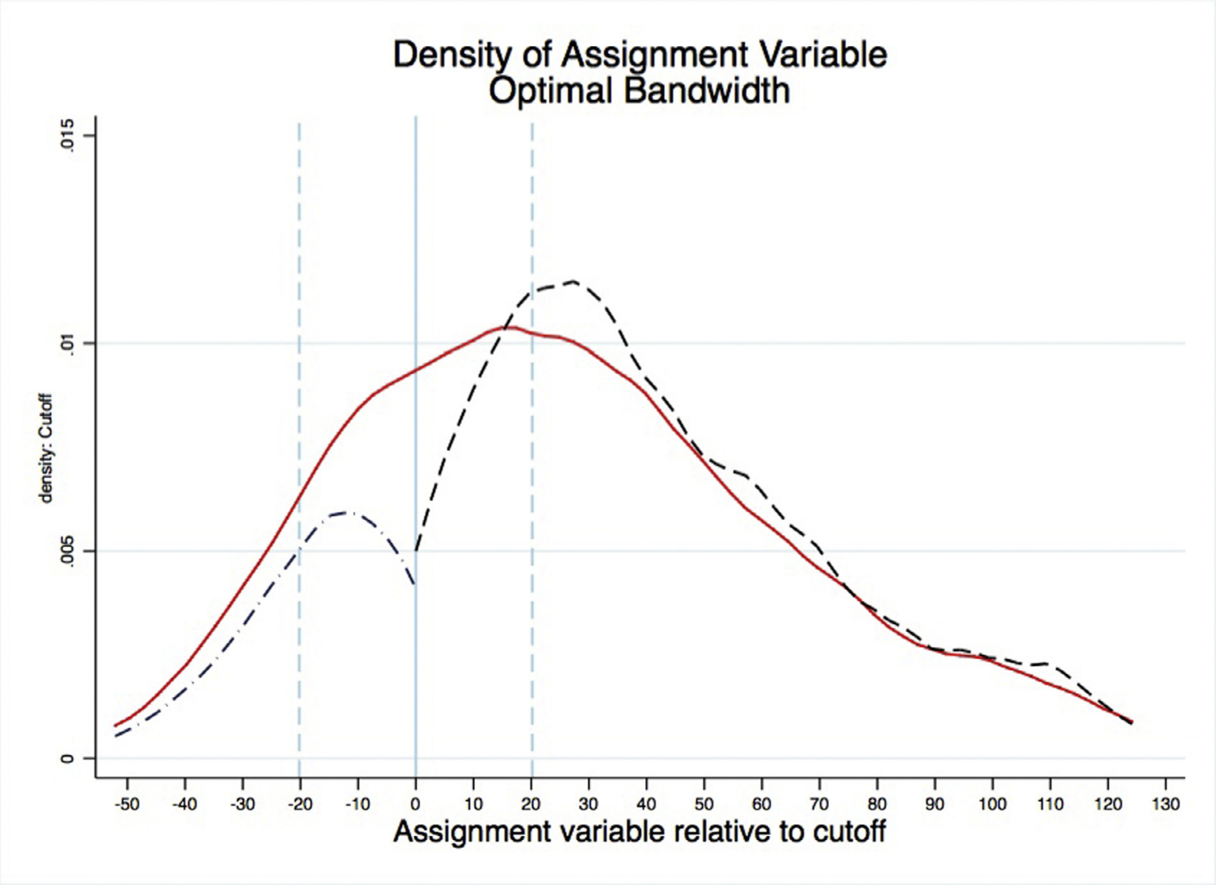


Fig. A2. Density of the assignment variable.

Table A1
Difference in difference PAE estimation, Spanish.

	(1)		(2)		(3)		(4)		(5)	
	All		All-controls		Optmal BW		Half the OB		Double OB	
PAE	−62.38***	(2.98)	−45.01***	(3.07)	−19.89***	(2.09)	−9.47***	(2.60)	−34.98***	(2.22)
2010	16.28***	(1.27)	16.74***	(1.31)	17.62***	(2.61)	20.25***	(3.80)	17.64***	(1.73)
2011	28.99***	(1.70)	29.80***	(1.72)	33.08***	(2.84)	33.84***	(4.43)	32.62***	(2.14)
2012	30.07***	(1.79)	31.05***	(1.83)	31.30***	(3.40)	32.92***	(4.17)	31.63***	(2.30)
PAE 2010	11.75***	(2.92)	8.65***	(3.03)	8.14**	(3.90)	7.81	(6.22)	8.26***	(3.03)
PAE 2011	14.56***	(3.16)	10.17***	(3.46)	8.07*	(4.13)	4.72	(5.66)	9.18***	(3.50)
PAE 2012	14.26***	(3.90)	9.40**	(4.08)	6.63	(5.08)	3.25	(6.15)	10.28**	(4.23)
Student/teacher			2.05***	(0.37)	0.52*	(0.30)	0.11	(0.46)	0.63**	(0.25)
Incentive program			20.45***	(5.45)	3.36	(4.65)	5.97	(5.69)	1.52	(3.94)
Teachers BA			−2.32	(7.37)	0.49	(6.07)	0.15	(8.31)	−1.40	(5.25)
Low marginality			−13.56***	(3.62)	3.44	(2.90)	2.02	(4.04)	2.85	(2.82)
Medium marginality			−19.96***	(5.27)	−1.32	(6.21)	−0.45	(6.72)	−7.64	(4.97)
High marginality			−15.50***	(5.71)	−4.07	(4.86)	−9.75**	(4.87)	−4.70	(5.38)
Constant	519.25***	(2.54)	455.07***	(10.82)	469.19***	(9.28)	474.83***	(12.07)	479.53***	(7.73)
R2	0.060		0.074		0.027		0.022		0.041	
Mean Dep										
SD Dep										
Obs	161,085		160,757		59,223		31,548		105,475	
Clusters	310		309		129		67		222	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

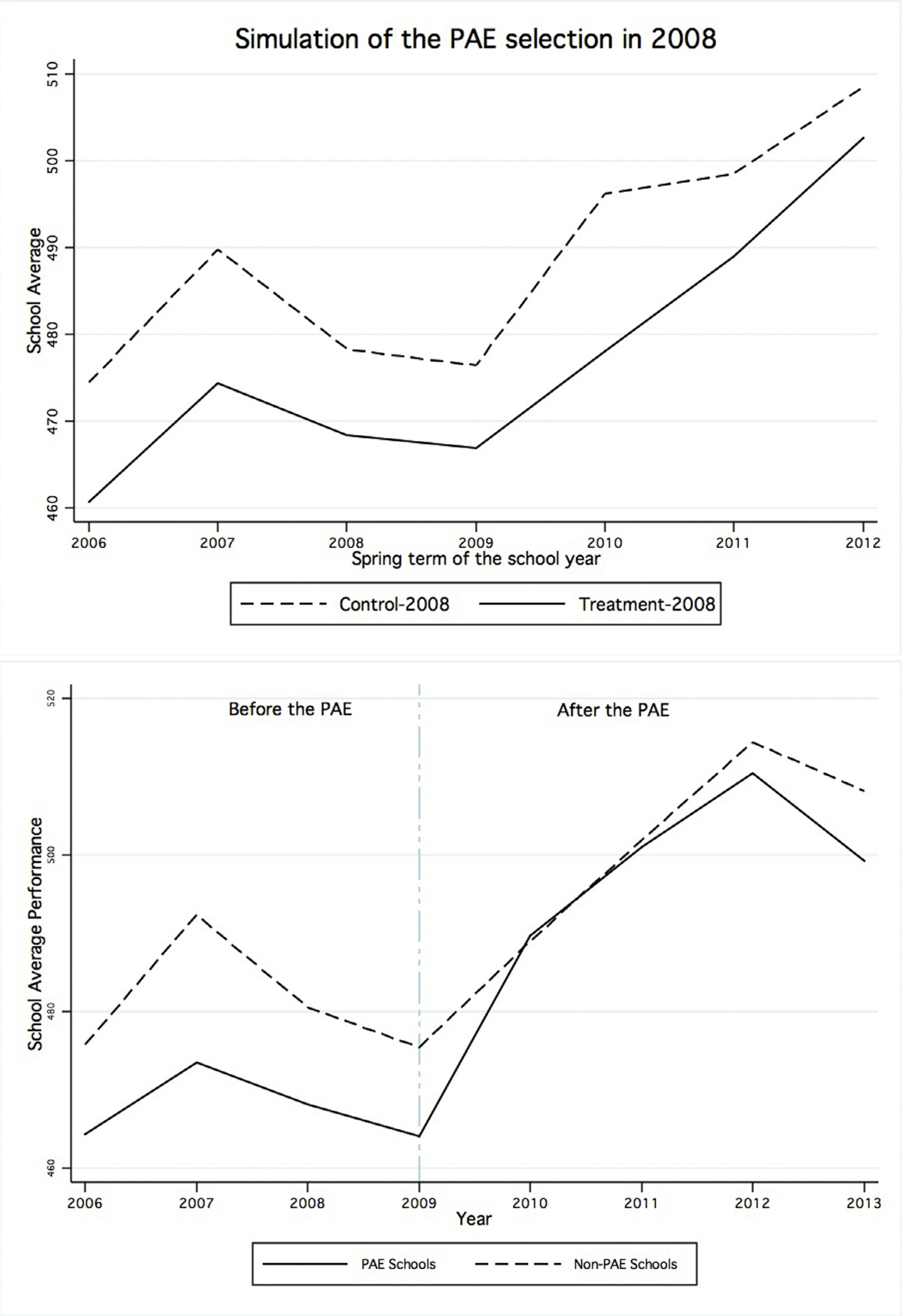


Fig. A3. Evolution of test scores under simulated (top) versus actual (bottom) PAE.

Table A2
Difference in difference, before vs. after estimation, math.

	(1)		(2)		(3)	
	All		All-controls		Optmal BW	
PAE	−64.71***	(3.15)	−47.86***	(3.31)	−22.55***	(2.68)
After	28.18***	(1.43)	29.02***	(1.41)	30.62***	(2.85)
PAE*after	18.27***	(2.94)	14.04***	(3.06)	12.61***	(4.04)
Student/teacher			2.00***	(0.39)	0.68**	(0.33)
Incentive program			22.95***	(6.11)	7.25	(5.11)
Teachers BA			−2.28	(7.98)	−9.09	(7.06)
Low marginality			−12.33***	(3.90)	5.26*	(3.06)
Medium marginality			−16.01**	(6.35)	5.12	(7.66)
High marginality			−15.62**	(7.09)	−1.91	(5.34)
Constant	523.28***	(2.69)	458.65***	(11.21)	473.02***	(9.68)
R2	0.052		0.064		0.026	
Obs	161,085		160,757		59,223	
Clusters	310		309		129	

Standard errors in parentheses.
Standard errors clustered by school.
* $p <$ 0.10, ** $p <$ 0.05, *** $p <$ 0.01.

Table A3
Difference in difference, before vs. after estimation, Spanish.

	(1)		(2)		(3)	
	All		All-controls		Optimal BW	
PAE	−62.38***	(2.98)	−45.07***	(3.07)	−19.96***	(2.10)
After	25.33***	(1.38)	25.98***	(1.38)	27.52***	(2.51)
PAE*after	13.55***	(2.65)	9.49***	(2.84)	7.63**	(3.52)
Student/teacher			2.06***	(0.37)	0.54*	(0.31)
Incentive program			20.05***	(5.47)	2.55	(4.66)
Teachers BA			−0.50	(7.36)	0.60	(6.06)
Low marginality			−13.46***	(3.64)	3.42	(2.91)
Medium marginality			−19.90***	(5.28)	−1.24	(6.22)
High marginality			−15.58***	(5.69)	−4.07	(4.90)
Constant	519.25***	(2.54)	453.49***	(10.75)	468.98***	(9.38)
R2	0.057		0.071		0.024	
Obs	161,085		160,757		59,223	
Clusters	310		309		129	

Standard errors in parentheses.
Standard errors clustered by school.
* $p <$ 0.10, ** $p <$ 0.05, *** $p <$ 0.01.

Table A4
Regresion discontinuity estimation results, Math 2011.

	(1)		(2)		(3)		(4)	
	All		Optmal BW		Half of OB		Double OB	
PAE	4.71	(6.43)	1.97	(8.06)	−2.72	(11.44)	8.90	(6.47)
Forcing variable	1.04***	(0.20)	0.66	(0.47)	1.31	(1.68)	1.04***	(0.22)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	−0.77	(0.60)	−0.12	(0.66)	−3.35	(2.14)	−0.08	(0.37)
$(PAE * (FV - cutoff))^2$	−0.01	(0.01)						
Student/teacher	0.43	(0.33)	0.14	(0.59)	−0.82	(0.77)	0.25	(0.38)
Incentive program	12.21**	(5.24)	14.68*	(8.62)	17.63	(11.34)	7.60	(6.16)
Teachers BA	−7.43	(9.03)	−8.78	(12.05)	0.50	(15.02)	−7.35	(9.84)
Low marginality	2.94	(3.68)	2.92	(4.70)	−3.66	(7.19)	4.84	(3.69)
Medium marginality	1.65	(6.60)	3.86	(11.01)	0.60	(13.38)	4.53	(7.43)
High marginality	−7.36	(7.12)	−7.95	(7.71)	−15.07**	(6.98)	−5.50	(7.90)
Constant	496.83***	(10.80)	507.54***	(16.08)	523.92***	(23.54)	502.17***	(12.25)
R2	0.074		0.007		0.007		0.023	
Mean Dep								
SD Dep								
Obs	39,866		14,635		7836		26,075	
Clusters	307		127		66		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p <$ 0.10, ** $p <$ 0.05, *** $p <$ 0.01.

Table A5
Regresion discontinuity estimation results, math 2012.

	(1)		(2)		(3)		(4)	
	All		Optmal BW		Half of OB		Double OB	
PAE	4.36	(7.82)	0.64	(8.39)	1.72	(10.71)	8.01	(7.13)
Forcing variable	1.01***	(0.21)	−0.19	(0.53)	1.02	(1.46)	1.07***	(0.19)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	−0.91	(0.93)	1.22	(0.81)	−1.76	(1.97)	−0.48	(0.43)
$(PAE * (FV - cutoff))^2$	−0.01	(0.02)						
Student/teacher	0.54	(0.40)	0.46	(0.61)	−0.07	(0.76)	0.72	(0.47)
Incentive program	14.04**	(7.11)	24.45**	(10.24)	39.68***	(13.01)	6.70	(8.56)
Teachers BA	−16.55	(10.76)	−10.29	(12.47)	0.09	(14.35)	−20.37	(12.35)
Low marginality	8.30*	(4.35)	13.57**	(5.50)	8.07	(7.40)	12.35***	(4.59)
Medium marginality	−1.29	(8.66)	−7.42	(9.26)	−8.90	(10.75)	0.97	(9.19)
High marginality	−7.57	(8.39)	−4.28	(8.90)	−13.32	(9.12)	−6.89	(7.69)
Constant	517.67***	(12.47)	517.73***	(17.05)	512.17***	(22.79)	516.60***	(14.69)
R2	0.063		0.012		0.014		0.020	
Mean Dep								
SD Dep								
Obs	43,806		16,261		8734		28,943	
Clusters	307		127		66		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6
Regresion discontinuity estimation results, Spanish 2010.

	(1)		(2)		(3)		(4)	
	All		Optmal BD		Half of OB		Double OB	
PAE	13.31*	(6.98)	6.63	(7.07)	11.81	(11.03)	9.70	(5.89)
Forcing variable	0.98***	(0.15)	0.63	(0.42)	1.85	(1.37)	1.01***	(0.16)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	0.99	(0.75)	0.69	(0.66)	−1.54	(2.13)	0.19	(0.30)
$(PAE * (FV - cutoff))^2$	0.03	(0.02)						
Student/teacher	1.11***	(0.30)	0.84*	(0.49)	1.69	(1.05)	1.19***	(0.39)
Incentive program	0.07	(4.13)	1.80	(6.03)	−2.38	(8.96)	−4.02	(4.66)
Teachers BA	−10.30*	(5.82)	−15.53	(10.31)	−49.82***	(14.11)	−11.67*	(6.61)
Low marginality	0.91	(2.97)	0.41	(3.95)	4.32	(5.95)	3.55	(2.99)
Medium marginality	−7.28	(5.96)	1.39	(9.82)	13.08	(10.42)	−3.29	(6.66)
High marginality	−1.21	(4.50)	−3.84	(3.75)	−1.93	(4.70)	−2.09	(4.26)
Constant	472.30***	(9.86)	485.74***	(16.72)	481.72***	(24.54)	471.71***	(12.53)
R2	0.102		0.010		0.012		0.035	
Mean Dep								
SD Dep								
Obs	38,928		14,201		7518		25,433	
Clusters	307		127		67		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7
Regresion discontinuity estimation results, Spanish 2011.

	(1)		(2)		(3)		(4)	
	All		Optmal BD		Half of OB		Double OB	
PAE	2.55	(5.81)	−0.20	(7.27)	1.58	(11.09)	6.01	(5.99)
Forcing variable	1.00***	(0.19)	0.76*	(0.44)	1.92	(1.81)	1.06***	(0.22)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	−0.72	(0.58)	−0.31	(0.61)	−2.48	(2.01)	−0.22	(0.35)
$(PAE * (FV - cutoff))^2$	−0.01	(0.01)						
Student/teacher	0.56*	(0.30)	0.14	(0.52)	−1.03	(0.69)	0.44	(0.35)
Incentive program	8.91*	(4.95)	6.59	(8.14)	13.32	(11.54)	1.97	(5.80)
Teachers BA	−7.98	(8.44)	−1.15	(10.91)	18.14	(13.95)	−4.04	(9.67)
Low marginality	2.45	(3.86)	3.55	(4.55)	−3.38	(7.29)	5.90	(3.91)
Medium marginality	−5.20	(5.36)	−3.45	(7.97)	−3.88	(10.26)	−0.98	(5.97)
High marginality	−6.01	(5.87)	−6.87	(7.29)	−15.27**	(7.41)	−3.37	(6.77)
Constant	495.92***	(10.73)	504.91***	(15.27)	515.69***	(22.02)	496.60***	(12.73)
R2	0.085		0.007		0.007		0.028	
Mean Dep								
SD Dep								
Obs	39,866		14,635		7836		26,075	
Clusters	307		127		66		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8
Regresion discontinuity estimation results, Spanish 2012.

	(1)		(2)		(3)		(4)	
	All		Optmal BD		Half of OB		Double OB	
PAE	1.67	(6.82)	−0.67	(7.85)	4.98	(9.88)	3.19	(6.38)
Forcing variable	0.98^{***}	(0.18)	0.34	(0.52)	1.94	(1.34)	1.00^{***}	(0.18)
FV square	−0.00	(0.00)						
$PAE * (FV - cutoff)$	−0.66	(0.77)	0.61	(0.75)	−1.81	(1.87)	−0.37	(0.36)
$(PAE * (FV - cutoff))^2$	−0.00	(0.02)						
Student/teacher	0.74^{**}	(0.33)	0.47	(0.61)	−0.22	(0.70)	0.88^{**}	(0.40)
Incentive program	11.10^{*}	(6.10)	8.47	(10.13)	19.30^{**}	(9.37)	0.84	(7.33)
Teachers BA	−6.11	(9.64)	1.84	(12.15)	12.07	(14.03)	−6.80	(11.01)
Low marginality	4.22	(4.06)	9.23^{*}	(5.41)	2.52	(6.63)	7.67^{*}	(4.14)
Medium marginality	−5.26	(6.99)	−7.94	(8.79)	−7.30	(10.39)	−2.71	(7.39)
High marginality	−6.31	(7.04)	−4.16	(9.70)	-17.78^{*}	(9.69)	−4.76	(7.65)
Constant	487.01^{***}	(11.03)	492.95^{***}	(17.24)	492.92^{***}	(19.04)	487.19^{***}	(13.11)
R2	0.084		0.009		0.008		0.024	
Mean Dep								
SD Dep								
Obs	43,806		16,261		8734		28,943	
Clusters	307		127		66		220	

Standard errors in parentheses.
Standard errors clustered by school.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Ahn, T., & Vigdor, J. (2014). The impact of no child left behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina. *Working Paper 20511*. National Bureau of Economic Research. doi:10.3386/w20511.

Andrabi, T., Das, J., & Khwaja, A. I. (2014). Report cards: The impact of providing school and child test scores on educational markets. *Working Paper Series rwp14-052*. Harvard University, John F. Kennedy School of Government.

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica, 74*(2), 431–497. doi:10.1111/j.1468-0262.2006.00668.x.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics, 119*(1), 249–275. doi:10.1162/003355304772839588.

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work : New evidence on accountability reforms*. The World Bank. Number 2270 in World Bank Publications.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis.. *Education Evaluation and Policy Analysis, 24*(4), 305–331.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review, 94*(4), 1237–1258.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9–10), 1045–1057.

Deke J, Dragoset L. (2012). Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials. Working Paper. Mathematica Policy Research, Inc., Jun.

Figlio, D. N., & Getzler, L. S. (2002). Accountability , ability and disability: Gaming the system. *NBER Working Papers 9307*. National Bureau of Economic Research, Inc.

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

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies, 79*(3), 933–959.

Imbens, G., & Lemieux, T. (2007). Regression discontinuity designs: A guide to practice. *Working Paper 13039*. National Bureau of Economic Research.

Imbens, G. M., & Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. *Working Paper 14251*. National Bureau of Economic Research.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*(3), 843–877.

Koning, P., & van der Wiel, K. (2013). Ranking the schools: How school-Quality information affects school choice in the Netherlands. *Journal of the European Economic Association, 11*(2), 466–493.

Loeb, S., & Strunk, K. (2007). Accountability and local control: Response to incentives with and without authority over resource generation and allocation. *Education Finance and Policy, 2*(1), 10–39.

Manacorda, M. (2012). The cost of grade retention. *The Review of Economics and Statistics, 94*(2), 596–606.

Mizala, A., Romaguera, P., & Urquiola, M. (2007). Socioeconomic status or noise? Tradeoffs in the generation of school quality information. *Journal of Development Economics, 84*(1), 61–75.

Mizala, A., & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics, 103*(C), 313–335.

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *Economic Journal, 120*(546), F187–F203.

Nichols, A. (2014). Rd: Stata module for regression discontinuity estimation,.

Reback, R. (2006). Teaching to the rating: School accountability and the distribution of student achievement. *Working Papers 0602*. Barnard College, Department of Economics.

Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy, 2*(4), 119–147. doi:10.1257/pol.2.4.119.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the florida heat? how low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251–281. doi:10.1257/pol.5.2.251.

Schochet, P. Z. (2009). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 34*(2), 238–266.