Economics of Education Review 66 (2018) 73–91

Contents lists available at ScienceDirect

Economics of Education Review

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

Concentrating efforts on low-performing schools: Impact estimates from a quasi-experimental design

Felipe Barrera-Osorio[a], Sandra García[b], Catherine Rodríguez[c,*], Fabio Sánchez[c], Mateo Arbeláez[d]

[a] Harvard Graduate School of Education, 13 Appian Way, Cambridge, MA 02138, United States
[b] School of Government, Universidad de los Andes, Bogota-Colombia, Cr 1 No 19-27 Bloque Aulas AU, tercer piso., Bogotá, Colombia
[c] Department of Economics, Universidad de los Andes, Bogotá-Colombia, Calle 19A No 1-37 Este. Bloque W, Bogotá, Colombia
[d] Ph.D. Student, Department of Economics, University of Illinois at Urbana-Champaign, 1407W Gregory Dr, Urbana, IL 61801, United States

ARTICLE INFO

Keywords:
Multi-level intervention
Regression discontinuity
Program design and implementation

JEL classification:
C29
C26
I21
I28
I29

ABSTRACT

This paper presents the impact evaluation results of the Colombian program *Todos a Aprender* (*Everyone Learning Program, ELP),* a multi-level intervention targeting low-performing schools. The main objective of the program was to increase math and language test scores of these schools through on-site teacher training, principal training and textbooks for students. Using census data from public schools containing detailed longitudinal information since 2010, the starting year of the program, and taking advantage of targeting rules based on dropout and grade repetition rates we fit a fuzzy regression discontinuity design to estimate program impacts. We also fit a difference-in-difference matching model as well as blocking with regressions to estimate the ATT impact of the program, based on observed characteristics used in the targeting process. Overall results indicate no significant impact of the program on test scores, grade repetition nor dropout rates. Additional analyses from a representative sample of 400 schools collected in the field suggest that deficiencies in the program's design and implementation could explain the lack of significant program impacts.

1. Introduction

The growing availability of international information on education shows two global trends in developing countries. First, most countries have achieved important advances in enrollment rates, particularly in primary education. Second, despite these advances, the overall gains in student learning, as measured by standardized test results, have been at best modest or null (Glewwe, Hanushek, Humpage, & Ravina, 2013; Pritchett, 2013). To meet the dual challenge of increasing access to education and improving learning outcomes, governments are implementing policies to tackle education from either the demand or supply side (Kremer, Brannen, & Glennerster, 2013; Murnane & Ganimian, 2014).

Most of the interventions addressing the supply side attempt to change specific aspects of a school; for example, by equipping them with computers or providing teacher training. On the other hand, a less common supply-side intervention – referred to as "multilevel interventions" (Snilstveit et al., 2015) – attempts to address more than one barrier for school quality improvement. These programs use a combination of interventions that in most cases include materials, teacher training, and infrastructure improvement. In some cases, these also include additional interventions, such as management training, school feeding, and diagnostic feedback. Evidence for the causal impacts of these type of interventions is limited. In a recent study, Snilstveit et al. (2015) identify only six studies that evaluated the impact of interventions that combined materials, teacher training, and management training in five low- and middle-income countries (Jamaica, Uruguay, Chile, Colombia and Mexico), finding mixed results on dropout rates and math and language test scores.

This paper contributes to the existing body of knowledge by presenting causal evidence of the impact of the *Everyone Learning Program* (ELP),[1] a multilevel intervention designed and implemented by Colombia's Ministry of National Education (MNE) since 2011. Likely constituting the most important program in the educational sector in the country in recent years, the basic goal of the program is to improve standardized test scores of primary education students attending public schools. The ELP, like most multilevel school interventions, focuses on the lowest performing schools in the country and aims to strengthen inputs in the educational process through the program's three main components: providing teacher training to improve teachers'

* Corresponding author.
E-mail address: cathrodr@uniandes.edu.co (C. Rodríguez).

[1] In Spanish, the name of the program is *Programa Todos a Aprender (PTA).*

https://doi.org/10.1016/j.econedurev.2018.07.001
Received 21 September 2017; Received in revised form 26 April 2018; Accepted 1 July 2018

pedagogical practices, providing training to principals to improve school management, and providing educational materials to students and teachers. Between 2011 and 2014, the ELP was implemented in more than 4,000 public schools across the country, reaching 77,086 teachers and nearly 1.9 million primary school students.

We use rich administrative data for the years 2009 to 2014 to evaluate the short-run impact of the program. By merging five administrative datasets, we construct a school panel data base that contains information for all public schools in the country. This information includes school characteristics such as location, number of students, number of teachers and schedules offered, as well as educational outcomes such as dropout and grade repetition rates, and the average test scores in math and language attained by their students in third and fifth grade. Additionally, we collected data from the field for this evaluation from a representative sample of 400 PTA-eligible schools in 2014. This data provides rich information about the actual implementation of the program, including the frequency of school visits and the type of activities performed as part of the intervention, and suggests the likely channels through which the results found in this paper can be explained.

To obtain a causal estimate of the ELP's impact, we exploit the discontinuity in the probability that a school would enter the program. The ELP focuses on the most underachieving schools, using four types of continuous variables to measure underachievement: the rate of grade repetition and student dropout, the variation in school enrollment, and standardized tests scores. By using these characteristics, we build an achievement index that allows us to employ a regression discontinuity design to evaluate the impact of the program on the probability of student dropout and grade repetition, and on student learning in the areas of language (Spanish) and mathematics around the cutoff point for program eligibility.

The results we obtain from the regression discontinuity design (RD) suggest that the program's impact on the quality and efficiency indicators in benefited schools is close to zero. These results are robust to different model specifications and cutoff points of program entry, as well as the use of different school samples. In general, the results do not allow us to reject the null hypothesis that the ELP has had no impact on rates of grade repetition and dropout, and on measures of educational quality obtained by students on standardized tests in mathematics and language.

The absence of a significant impact under the RD design could be attributed to the local character of the estimator. It is possible that, under heterogeneous effects, schools near the cutoff point for program entry fail to present evidence for program impacts. It is also possible to argue that treated schools further from the cutoff point, which would be the most underperforming schools, have benefited from the ELP. To test this hypothesis, using a larger sample, we estimate the average impact on the treated schools under difference-in-difference matching models as well as matching under blocking with regressions (Imbens, 2015). The results are consistent with those obtained from the RD design. On average, the DD-matching and blocking models show no impact of the program on dropout, grade repetition, or standardized test scores for this broader sample of schools.

One plausible explanation for these (null) results is that not enough time has elapsed to observe effects in educational quality and efficiency rates, and any impact evaluation of the program needs to be executed later in time. Alternatively, deficiencies in the program design or implementation could explain, in part, our results. Additional analyses carried out with information from a complementary representative sample of 400 eligible schools collected in the field suggest that, although the program has impacted some short-term variables related to teaching practices, deficiencies in the program design and implementation might be the main explanation behind the lack of significant program effects. We find that the number of yearly individual meetings of beneficiary teachers with their tutors was very low and far below what was originally planned. The mean number of visit per year for teachers in our sample was 3, while the programme foresaw one visit per week. Furthermore, in these few sessions, no evidence of a clear structure of core activities undertaken is present.

The contribution of this study to existing knowledge about the causal effects of multilevel interventions is twofold. Firstly, through the causal identification strategy, we show that the positive effects previously found for similar programs in Chile (Chay, McEwan, & Urquiola, 2005) and Uruguay (Cerdan-Infantes & Vermeersch, 2007), do not extrapolate to other contexts. The evidence suggests that ELP has had no short-term impact on educational outcomes of the benefited students. Secondly, the evidence also highlights the importance that appropriate program design and implementation can have. The program fell short on clear strategies and program design, which explains in part the results of this paper and constitutes important feedback for policy implementers.

The remainder of the paper is organized as follows. Section 2 summarizes the main results from the literature, while Section 3 describes the context, goals, and implementation of the ELP. Sections 4 and 5 describe the data and identification strategies we use in this evaluation. We present our main results in Section 6 and explore the possible channels that could explain these results in Section 7 . Finally, we present our conclusions in Section 8 .

2. Previous studies

To induce higher enrollment rates and increase the education quality across the public education system, several countries are currently implementing multilevel interventions that aim at improving various schooling inputs. Even though multilevel interventions are not homogenous across countries, they share three common features: (1) most programs focus on low-performing schools; (2) all programs combine at least two interventions at the school and teacher level, and; (3) most programs include the provision of materials and teacher training in combination with infrastructure rehabilitation or leadership training. It is important to underscore that there are other types of interventions at the school level that have more than one component, but that are not necessarily multilevel interventions as they are only targeted to one organizational level. For instance, school management interventions such as the School Management Program in Brazil (Tavares, 2015) include training to school managers (principals and coordinators) and development of monitoring indicators and action plans, but do not include a teacher-level component.

The most recent review conducted by Snilstveit et al. (2015) shows 10 studies reporting causal evidence of multilevel interventions in low and middle-income countries. The authors conduct a meta-analysis showing that multilevel interventions have, on average, improved standardized math test scores by 0.16 standard deviations and language test scores by 0.04 standard deviations. However, they also report large heterogeneity in effect sizes, ranging from a negative impact for third grade students in China and for urban students in Mexico, to modest positive effects for primary students in Chile and large positive effects for fifth grade students in China (Snilstveit et al., 2015).

Among the 10 studies that examine the effect of multilevel interventions on educational outcomes, five programs share at least three common characteristics with the ELP: the combination of materials, teacher training, and management training at the school level. Also, one program has in common the combination of materials and teacher training, but the management piece corresponds to governance at the subnational level (“Rural Education Project” in Colombia). The evidence for the impacts of these programs on language test scores is mixed, with positive effects in the case of Chile (Bellei, 2011; Chay et al., 2005), Colombia (Rodríguez, Sánchez, & Armenta, 2010), Uruguay (Cerdan-Infantes & Vermeersch, 2007), and rural Mexico (Paqueo & Lopez-Acevedo, 2003); no effects in the case of Jamaica (Lockheed, Harris, & Jayasundera, 2010); and negative effects in the case of Mexico for urban students (Paqueo & Lopez-Acevedo, 2003).

Evidence on the impacts of these programs on math test scores is scarce but consistent: studies for programs both in Chile and Uruguay report positive effects (Bellei, 2011; Cerdan-Infantes & Vermeersch, 2007; Chay et al., 2005). Among the studies reviewed by Snilstveit et al. (2015) that share characteristics with the ELP, only the Colombian program "Rural Education Project" has evidence for efficiency measures, with significant impacts on the reduction in dropout and failure rates (Rodríguez et al., 2010).

It is important to note that some of these multilevel interventions, in addition to materials, teacher training, and management training, include interventions such as school feeding and parenting programs (in the case of Jamaica), monetary teacher incentives (in the case of Mexico), or lengthening the school day (in the case of Uruguay). Therefore, the impacts reported cannot be exclusively attributed to the bundle of materials and teacher training.

The Chilean experience seems to be the closest to the Colombian context. In 1988, Chile implemented a program known as *P-900*, which sought to improve the quality of education through infrastructure enhancement, the provision of supplementary instructional materials, teacher training, management training, and additional classes for the lowest-performing students provided outside of normal school hours. A causal evaluation of the program, using an RD design, shows positive but modest results (Chay et al., 2005). Years later in the same country, the "*Technical Support to Failing Schools Program,*" implemented between 2002 and 2005, sought as its primary objective to increase the standardized test scores of first through fourth grade students. As with the ELP, each benefited Chilean school received instructional materials and was visited by an external advisor to improve classroom practices and school administration. Using a RD methodology, Bellei (2011) finds that the program had positive effects in the short run on students' scores for language (Spanish) and math; nonetheless long-term effects are only found in language test scores.

3. The ELP program context

Colombia has followed the international education trends mentioned in the introduction. Like most developing countries, although enrollment rates for primary education level have significantly increased in recent years, an important number of children and adolescents still do not attend pre-school or complete secondary education. Similarly, while important efforts and programs have been implemented to improve the quality of education in recent years, these have been insufficient and unsystematic. Both national and international assessments have shown that the quality of education in Colombia is low and extremely unequal (Barrera, Maldonado, & Rodríguez, 2014).

To overcome the shortcomings on the quality of the education received by the most poorly performing students, the Colombian government launched the Everyone Learning Program (ELP) in 2011. The ELP, which probably constitutes the most important program in the educational sector in recent years, is a comprehensive strategy with the objective of improving the learning conditions within a specific group of low-performing public schools in the country. Specifically, the original goal was to bring the program to 3000 schools, raising the standardized test scores of at least 25% of these schools' students, in the areas of language (Spanish) and mathematics, as measured by the Saber 3 and 5 standardized exams taken by all third and fifth grade students in 2014, respectively.[2] The program sought to benefit 2,300,000 primary students in language and math, while training and advising 70,000 educators, including teachers and school principals.

To reach these goals, the ELP offered a combination of three complementary strategies, including (1) the provision of educational materials and diagnostic exams for students, (2) teacher on-site training, and (3) support to principals.[3,4] The first component aimed to meet competency standards by providing adequate materials, using student evaluations, and improving pedagogical practices. To this end, the program provided textbooks in language and mathematics to all primary students (grades 1 through 5). Additionally, a diagnostic exam provided individual results for each student at least once per year, giving teachers feedback on the strengths and weaknesses of their students, including which areas the teacher should focus on.

The second component (teacher on-site training) is arguably the most important in terms of the monetary and time resources that were invested. This component sought to improve teaching practices through the direct observation and mentorship of teachers by tutors with teaching experience, selected for the program based on merit.[5] The original design of the program aimed to bring tutors to schools for several class sessions to provide support to teachers in daily activities, such as planning, evaluating, and designing appropriate teaching strategies, to achieve maximum impact and, consequently, maximize students' learning. According to the program design, once a school was selected as an ELP school, all elementary teachers were invited to participate. Although teachers were not obliged to participate in the program, school principals and coordinators encouraged participation of all primary teachers.[6] Tutors visited schools and were expected to have weekly group meetings with participating teachers as well as individual sessions that included class observations and feedback on teaching practices (as discussed later in the paper, in practice the frequency of meetings was lower than originally planned). Group meetings were the same for all teachers in all schools, while individual sessions were supposed to be teacher-specific depending on the needs observed by each tutor in their trainees.

In general, group meetings took place after school hours and individual meetings took place during school hours; therefore part of the training took place during teachers' working hours. Teachers were not paid for the extra time devoted to the portion of the training that took place outside their working hours. While teachers' attendance was not checked on an individual basis, tutors were required to report to the program manager (at the Ministry of Education) the number of participant teachers on a regular basis. Also, they were required to upload the working plan set up with teachers, including objectives and action plans, to a centralized information system.

[2] Source: MNE. http://www.mineducacion.gov.co/1621/w3-article-304244.html.

[3] Initially, an additional component was to be included to improve the basic conditions of these schools (infrastructure, transportation support, and school nutrition programs), but this part of the program lacked sufficient resources and had an insufficiently clear intervention strategy.

[4] It is important to mention the ELP program is not the first on-site teacher training program implemented in Colombia. Between 1999 until 2008 the Colombian government implemented the first phase of the Rural Education Program (PER for its acronym in Spanish). During this period, this supply-side scheme program implemented flexible educational models adapted to the needs of the rural communities in the country through the provision of specialized didactic material and on-site teacher training. As above mentioned, Rodríguez et al. (2010) show the program had significant impacts on measures of efficiency (dropout, passing, and failure rates) and on education quality.

[5] Tutors were selected by the Ministry of Education on a merit basis from a pool of candidates that applied to a public call for teachers in the public and private sector. Tutors who were teaching in non-ELP public schools at the time of application were temporarily transferred to ELP schools (as tutors) and were not assigned any classroom activities with students. Most of the time tutors were assigned to the same school district (zone) but not to the same school where he/she taught.

[6] Primary teachers in Colombia oversee a single classroom and teach all basic areas of knowledge to their students including Spanish, Mathematics, Science, and Social studies. They are not obliged to have any specific training in each area of knowledge, but only a professional degree. Thus, teachers that were trained by ELP have different undergraduate studies and, unlike tutors, were not restricted to a math or language higher education diploma.

Finally, the third component targeted organizational improvement by training school principals and providing information about sources of monetary resources to improve school conditions. According to the program design, training to school principals consisted of two main activities: access to a school management course provided by a public university and on-site sessions with the tutor. These sessions were mainly devoted to an assessment of the Institutional Improvement Plan, including monitoring of goals and revision of strategies in order to achieve quality goals identified in the improvement plan.

The selection of schools for entry into the ELP was carried out in three stages: the initial selection of educational institutions by the MNE (Stage 1); the selection of additional educational institutions by consulting firm McKinsey & Company (McKinsey), which was at that moment contracted by the Colombian government for this purpose (Stage 2); and the self-selection of some schools in Certified Territorial Entities (CTEs) that wished to participate in the program and, although they did not meet the initial criteria, contributed resources for the program's implementation (Stage 3).[7]

During Stage 1 of the selection process, the MNE's first criteria to select beneficiary schools was to focus on those located in previously-identified priority CTEs requiring critical interventions. Specifically, the MNE has a color-coded prioritization system consisting of categorizing CTEs according to their needs, with the purpose of designing the types of interventions they require. There are three groups: red (critical situations requiring comprehensive interventions), orange (difficulties in specific areas requiring more directed program targeting –, such as quality/coverage, early childhood, or management), and green (acceptable, stable conditions).[8] For the specific case of the ELP, schools selected in Stage 1 all belonged all to CTEs classified with either the red or orange color. These CTEs are characterized by low enrollment rates and/or inadequate quality of education. Moreover, many of them serve students from low socioeconomic backgrounds. Then, within the group of schools located in these critical areas, the MNE created a selection rule under two additional criteria associated with efficiency and quality indicators. The first criterion was fulfilled if the grade repetition rate of students in the school was over 4.6% in 2010, or if the school was categorized as a low-achieving school in the Saber 5 standardized test results of that same year.[9] The second criterion was fulfilled if the school had a dropout rate above 3.8% in 2010, or a decrease in school enrollment between 2009 and 2010. In total, 1,386 schools were selected to receive treatment in the first selection phase of the program.

In Stage 2 of the selection process, the MNE gave the list of preselected schools to McKinsey & Company, which added new schools with the goal of optimizing program effects. The selection during this second stage sought to raise the potential impact of the ELP in terms of student coverage and cost-efficiency indicators. To this end, two additional selection criteria were added, favoring institutions with larger enrollment and easier geographical access. Using these criteria, an additional 990 schools were selected to benefit from the program in this second stage.

Finally, in Stage 3 of the selection process some CTEs petitioned the MNE for the inclusion of certain schools into the program. Some of these CTEs even offered to cover part of the cost of the intervention in these schools. This allowed the MNE to offer the program to schools that had initially been identified as potentially eligible for the program, given their low achievement. In total, 1904 schools were included in this group, 90% of which were categorized as schools with minimum achievement or as low achievement/priority cases, but which did not necessarily fulfill the criteria of high grade repetition or school dropout rates.

Table 1 summarizes the final number of selected schools in each of the three original stages according to the color of CTE they belonged to. It also has information on the schools not selected for the program some of which, depending on the estimation strategy, will serve as our control schools.[10] In total, at the end of the three selection stages, 4890 public schools were selected for treatment, representing nearly 30% of Colombia's public education institutions. Of these, by 2014, the ELP had effectively reached 4255 through a gradual expansion over time. The first year of national coverage was 2012, with 1829 schools starting treatment early in that year. By 2013, the ELP program had already reached most of the selected schools, for a total of 4071. Finally, in 2014 this number increased to 4200 schools. Thus, by the end of that academic year, most schools had experienced between 2 or 2.5 years of exposure to the ELP. It is the impact of this exposure that we evaluate in this paper.

4. Data

This study uses five different datasets to evaluate the impact of the ELP on treated schools. The first four are drawn from administrative information collected in countrywide censuses, while the fifth comes from field data collected from 400 schools for this impact evaluation in 2014.

The first database uses information from Colombia's Integrated System of School Enrollment (SIMAT for its Spanish acronym) which compiles individual-level data for all public-school students in Colombia between 2006 and 2015. This database contains the educational history of every student, including the specific school they attend each year, the grade they are in, individual national IDs, names, and date of birth. This information allows us to construct a longitudinal dataset at the school level, which we then use to calculate dropout and grade repetition rates for each school for each year.

The second information source is the Colombian Institute for the Evaluation of Education (ICFES, for its Spanish acronym), from which we obtain the scores of the Saber 5 standardized exams for 2009 and 2014 and for Saber 3 for the year 2014, as the exam for this grade was introduced after the ELP program was in place. Saber is a are national standardized exams taken by all third and fifth grade primary school students in the areas of mathematics and language, and reported at the school level by the ICFES.

The third source is the C-600 national school census survey collected by the National Administrative Department of Statistics (DANE for its Spanish acronym), which provides general information on schools for years 2009 and 2010 (pre-treatment characteristics) and was used by the MNE in the school selection process for program treatment. We draw information from the C-600 including: the number of students in each school; the number of teachers in each school; the

[7] In Colombia, the CTEs are the departments (similar to states in the USA), districts, and municipalities with more than 100,000 residents. These CTEs have the autonomy to manage educational resources and organization within their territories, within the established parameters required by Colombian law. Currently, the education sector in Colombia has 94 CTEs: 32 departments, 4 districts, and 58 municipalities.

[8] A detailed summary of the system's design can be found in Barrera, Maldonado, and Rodríguez (2014).

[9] The classification of low achieving schools was done by the MNE based on the Saber 5 results of 2009. However, we do not have the details of the criteria used to do decide this. The data base available to the researchers simply has information of whether a school was or not designated as one with quality problems.

[10] Although not shown in the Table, there were 116 schools that were benefited by ELP, even though were not originally selected for treatment in any of the three stages described. Given that we have no information on their selection process, we omit these schools from the analysis. Similarly, there was a group of schools that only received books, due to a surplus in the program's available resources (the provider of materials was chosen in a bidding process and the winning bid was substantially below the MNE's original budget, resulting in a surplus of materials). For our analysis, the treatment variable is only assigned to schools that received both mentoring and materials; schools that received only materials are left out of the analysis as well.

Table 1
Selection of schools in each stage of ELP.
Source: MNE, authors' calculations.

Selection stage	Type of ETC they belong to				
	Green	Orange	Red	Not classified	Total
Stage 1	0 (0%)	408 (25.6%)	1,188 (74.4%)	0 (0%)	1596
Stage 2	383 (27.4%)	619 (44.3%)	395 (28.3%)	0 (0%)	1397
Stage 3	88 (4.6%)	1161 (61.1%)	605 (31.8%)	46 (2.4%)	1900
Not selected for ELP	934 (7.5%)	7015 (56.0%)	1224 (9.8%)	3349 (26.7%)	12,522
Total	1405 (8.1%)	9205 (52.9%)	3413 (19.6%)	3395 (19.5%)	17,415

level of education of teachers in each school; whether the school offered a morning, afternoon or double shift; and if the school is in a rural or urban area.

The fourth source comes from the ELP program's administrative data, including information on the schools that received treatment, the specific dates when schools were visited by the tutors, and other general characteristics on the program's implementation such as when they received the specialized materials. It is worth highlighting that the four databases use a unique ID to identify each school, allowing us to merge all information into a single database to conduct the analyses for this paper at the school level.

The fifth and final source of information we use comes from a sample of 400 schools chosen for the evaluation of the ELP's impact on teaching practices. School selection for this sample was based on a matching methodology in which 200 treated and 200 untreated schools were randomly chosen and matched based on most-similar characteristics as possible (García, Harker, Figueroa, Gómez-Echeverry, & Rojas, 2017). The selection sought to create a representative sample of all benefited schools. Between August and October of 2014, we visited the schools, and applied specially designed surveys to principals, teachers, and students. We also conducted detailed classroom observations to both treatment and control teachers. These survey instruments provide evidence on the quality of the implementation of ELP in the representative sample of treated schools and its information sheds light on possible transmission mechanisms that may explain the results obtained in this impact evaluation.[11]

Table 2 presents a comparison of the main pre-treatment characteristics between ELP treated schools and other public schools not benefited by the program. The efficiency and quality indicators of schooling outcomes show, as expected, that on average schools selected for the ELP had higher dropout rates and lower average scores on the Saber exams in 2009. Considering the selection criteria used in Stage 2 of the selection process, we can also see that the percentage of institutions located in rural areas and the percentage of the student population considered "vulnerable" is lower in treated schools. Treated schools have, on average, a larger number of teachers and students than non-treated schools.

These pre-treatment differences, all statistically significant, suggest that a causal evaluation of the potential program's impacts must be carefully designed and implemented. In fact, Table A.1 shows that, as expected, even when comparing treated schools entering in different selection stages significant differences emerge. The institutions initially selected by the MNE and those selected by the CTEs in Stage 3 exhibit higher levels of vulnerability than those selected by McKinsey in Stage 2. For example, schools selected in the first stage have higher dropout and grade repetition rates, and lower aggregate levels of achievement in mathematics and language, than those selected in the following stages. Moreover, schools selected by McKinsey in Stage 2 are slightly better off than those selected in the first and third stages in terms of socioeconomic status and enrollment. On average, the schools selected by McKinsey have a lower percentage of students living in vulnerable conditions, are more likely to be in urban areas, and have more students and teachers.

5. Identification strategy

The description of the context in which the ELP's was implemented and the analysis of the descriptive statistics presented in the previous sections indicate that the selection process for program entry was not random and hence treated schools differ in many ways from the non-treated. To overcome the implications of the selection process and obtain the causal impact of the program on efficiency and quality indicators we use three complementary identification strategies: (i) a regression discontinuity (RD) design for the schools selected in Stage 1 by the MNE, (ii) a difference-in-difference matching approach for all ELP-treated schools; and, (iii) blocking estimations for all ELP-treated schools. Below, we discuss each of these strategies in detail.

5.1. Regression discontinuity design

To estimate the causal impact of the ELP on efficiency and quality indicators through a RD design, we take advantage of the program selection mechanism, which was clearly defined and difficult to manipulate in the first stage. Following the selection steps in this stage, we took as our sample of interest only those schools that fulfilled three main criteria: (i) the school was classified in one of the two lowest socioeconomic categories by the MNE; (ii) the school was in a CTE classified in a category where ELP schools from Stage 1 were also located (that is, CTEs with serious quality and coverage problems classified either as red or orange); and, (iii) the school's results on the 2009 Saber standardized exams placed it in the low achievement category. All these were criteria that, as mentioned, were considered by the MNE in selecting the beneficiary schools in the first stage. Table A.2 details the schools that were eliminated from the RD sample depending on the different selection criteria and whether they belong to the treatment or the control group. From the former group, only 30% of treated schools are kept: the ones that were selected in the first step by the MNE. From the control group, we are left with 35% of all schools in the original panel dataset. Most control schools were eliminated, as they did not belong to a CTE where a treatment school was located or because they did not have complete efficiency information. Thus, the final RD sample contains information for 2623 schools on which the local impact of the ELP program will be evaluated through an RD design.

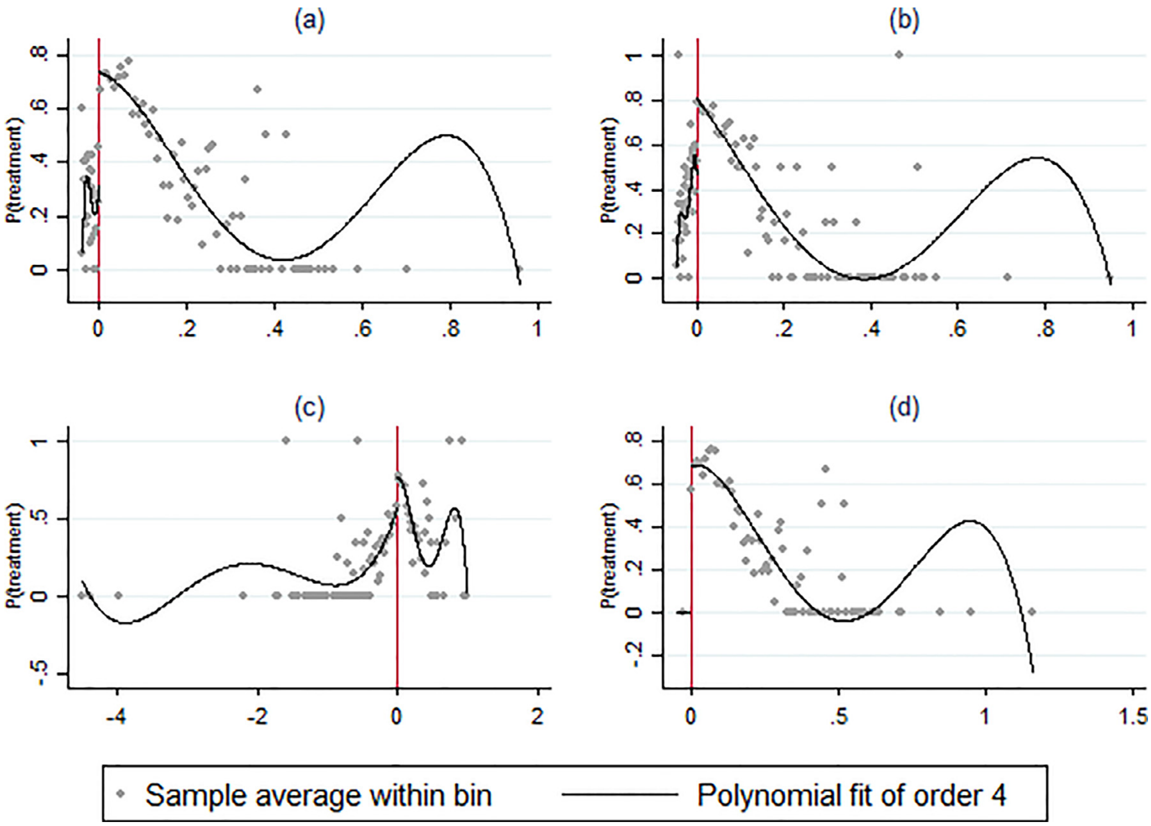
For this specific RD sample of schools, we analyze how the probability of treatment by the ELP changes according to the three efficiency measures defined by the MNE: dropout rates, grade repetition rates, and variation in enrollment in 2010. As previously described, the MNE defined as selection criteria for program entry a minimum 4.6% grade repetition rate in 2010, a minimum of 3.8% dropout rate in 2010, or a decrease in school enrollment between 2009 and 2010. The first three panels of Graph 1 present the treatment probability according to these three criteria, respectively, where the red continuous line represents the cutoff values for each criterion. A visual inspection of the treatment probability and the three measures of efficiency shows a discontinuity in the probability of selection occurring precisely at the cutoff thresholds defined by the MNE, especially for the first two measures related to dropout and grade repetition. The discontinuity is close to 40 percentage points when the dropout rate is used as the running variable and decreases to nearly 30 and 20 percentage points when change in enrollment and repetition rates are instead used respectively.

[11] Results of the comparison between treatment and control samples on teaching practices are reported by García et al. (2017). Here we report results on implementation variables from the treated subsample.

Table 2
Descriptive statistics of pre-treatment characteristics for treated and non-treated ELP schools.

Variables considered for program participation	Treatment group			Control group			Difference (*t*-test)
	Mean	SD	N	Mean	SD	N	
Variation in enrollment 2009–2010 (assignment variable)	0.011	0.404	4,201	0.082	1.999	9,138	−0.071**
Dropout rate (assignment variable)	0.073	0.057	4,006	0.067	0.087	9,222	0.007***
Grade repetition rate (assignment variable)	0.057	0.053	4,006	0.072	0.098	9,222	−0.014***
Saber5 2009 – Math average	274.700	39.1	3910	306.1	45.7	4435	−31.317***
Saber5 2009 – language average	274.200	32	3889	298.8	37.9	4394	−24.637***
Vulnerable population	0.684	0.2	4191	0.748	0.28	9237	−0.064***
Rural schools	0.587	0.492	4255	0.772	0.419	10,731	−0.185***
Number of teachers	24.000	20.7	4037	9.319	17.75	11,746	14.716***
Enrollment 2010 R166	1059.000	987.6	3916	396.5	5111	10,960	663.012***
Full-day Schedule	0.230	0.421	4204	0.422	0.494	9473	−0.192***
Morning Schedule	0.915	0.279	4,204	0.658	0.474	9473	0.257***
Afternoon Schedule	0.582	0.493	4204	0.28	0.449	9473	0.301***
Evening Schedule	0.459	0.498	4204	0.265	0.441	9473	0.195***
Weekend schedule	0.556	0.497	4204	0.288	0.453	9473	0.268***

*** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.



Graph 1. ELP treatment probability, given dropout rate, grade repetition rate, enrollment variation, and maximization index for schools in 2010.
Notes: Each panel presents the probability of treatment for a school given its rates of dropout (a), grade repetition (b), enrollment variation (c), and maximization index (d) in 2010 respectively. Source: MNE, authors' calculations.

However, the graph also shows that none of the three defined thresholds in Stage 1 perfectly determines treatment. This is not surprising given that the MNE selected any school that met at least one of these three criteria, provided that the low achievement criterion was also met. That is, a school might have been selected if its students achieved a low-test score average on the Saber exams and only one of these three criteria were above the thresholds. To maximize the difference in the probability of treatment, following Papay, Willett, and Murnane (2011), we created an index composed of the grade repetition and dropout rate selection variables that closely followed the MNE collective selection criteria. The index is estimated as:[12]

$$\text{Maximization Index} = \max\{drop, rep\}$$

[12] We also attempted to create this index including variation in enrollment. Specifically, we estimated the index as max{varia,des,rep} where varia is the enrollment variation. The conditions required for a strong first stage in the fuzzy RD design were not as strong and thus we use the index without including this variable.

Where *drop* and *rep* are the normalized dropout and grade repetition rates adjusted to a single maximum value that identifies if a given public school had either the dropout or the repetition rates above the cutoff point assigned by the MNE in the first selection process.[13]

Panel d of Graph 1 shows that the discontinuity in the probability of receiving treatment is nearly 60 percentage points when using this index as our selection criterion. Treatment probability is zero to the left of the index if the school does not comply with any of the two criteria used in our index and increases to almost 60% just to the right of it. This increase is unsurprising, as the index allows us to capture the fact that selection occurs if any of the two criteria is met, either grade repetition or dropout rates above the MEN's selected threshold, though not necessarily one criterion in particular.

Using the four alternative running variables, we fit a fuzzy regression discontinuity design to find the causal impact of the ELP program. We use the "rdrobust" command that implements the bias-corrected inference procedure proposed by Cattaneo, Calonico, and Titiunik (2014). This is a data-driven local-polynomial-based inference procedure robust to large bandwidth choices. We estimate the local impact of ELP on the following academic outcomes: primary grade repetition rates of students in school *i* in year *t;* the primary dropout rate of students in school *i* in year *t*, and the measures of educational quality and learning achievement for these students, as measured by the Saber 3 and 5 standardized exams in 2014. Under this setting we obtain a causal RD-IV estimand of the form:

$$\theta_{RD-IV} = E\left[Y^{1}_{i,2014} - Y^{0}_{i,2014} | R_i = k,\;\; D^{1}_{i} \rangle D^{0}_{i}\right] \quad (1)$$

Where $Y^T_{i,t}$is the outcome variable of interest above mentioned for school *i* in time *t* (2014) for treated and control schools respectively (T = 1 or T = 0); R_i is one of the four running variables (grade repetition, dropout rate, change in enrollment, and our constructed index); *k* is the point of discontinuous assignment rule for each running variable; and D^T_i an indicator variable if school *i* is benefited by the ELP program.[14]

Besides discontinuity in treatment probabilities, another necessary assumption to obtain a valid RD estimator is that there is no manipulation in assignment to treatment around the cutoff point for program eligibility. There are several arguments in favor of this assumption in the ELP case. First, the dropout and grade repetition rates used for program selection were estimated by the central administration before the existence of the ELP program. Neither a school's principal nor its teachers knew the program would ever be implemented, much less the cutoff points that would be used for program selection. Secondly, the final index used in our identification strategy was not known by the schools nor the MNE at any point in time. Nonetheless, Graph 2 provides further statistical evidence of the absence of any manipulation of treatment status around the cutoff by presenting the distribution of the four alternative running variables used as the eligibility criterion as suggested by McCrary (2008). For all four variables there is no significant evidence of discontinuity around any of the cut-off points – particularly for our index – confirming that there is no statistical evidence of systematic manipulation of the running variable.

Finally, a valid RD design requires that the schools falling closely to the left of the cutoff point are similar to those falling closely to the right of the cutoff (the treated group), so that the only difference between these two groups is the probability of treatment. As such, it is essential to demonstrate that there are no significant differences in the pre-treatment measures of our control variables in schools located close to either side of the cutoff point. Using again the four alternative running variables, Table A.3 presents the results of the regressions having pre-treatment observable covariates as the outcome variables of interest. As observed, the vast majority of the efficiency and quality education outcomes of schools near each of the four cutoff points are not statistically different between eligible and non-eligible schools before the program was implemented, providing effective evidence that our RD strategy is appropriate. Nonetheless, as suggested by Cattaneo, Jansson, and Ma (2016a,b), the estimations control for those pre-treatment variables where a significant difference was found.

5.2. Difference-in-Difference matching and blocking

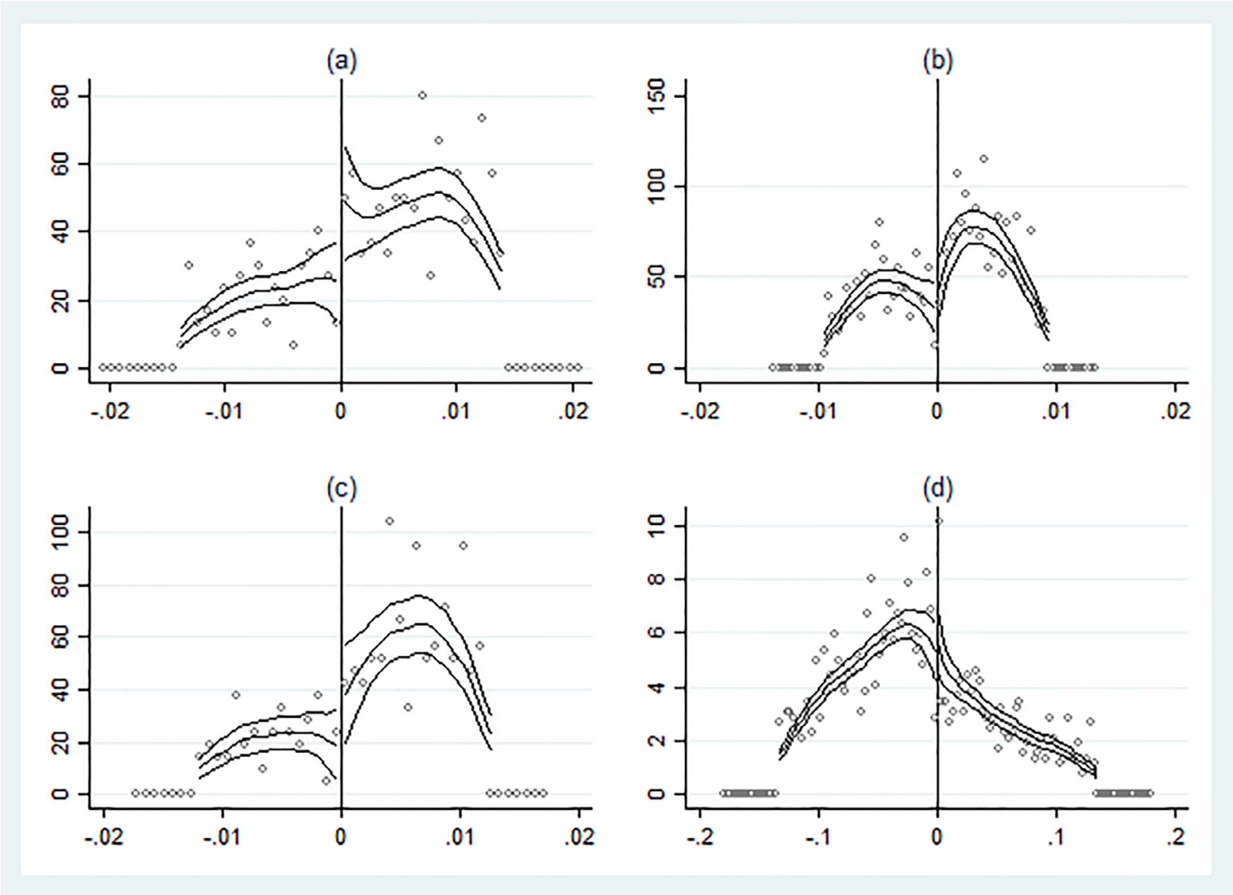
Although the estimands obtained through the RD estimations provide a quasi-experimental causal impact of the ELP, the greatest disadvantage of this strategy is that we can only interpret these results as a local effect of the program. Moreover, it only presents the impact on those schools that are close to the cutoff point and were thus selected in the first stage of the ELP selection process, representing only 30% of all treated schools. To complement the analysis and provide more general estimations of the ELP's impact for the complete group of treated schools, we also use a difference-in-difference matching methodology, as proposed by Heckman, Ichimura, and Todd (1997). This methodology combines the characteristics of non-parametric methods from traditional propensity score matching estimators, and limits self-selection problems based on constant non-observable characteristics, as in the DID method. These authors demonstrate that with databases sufficiently rich in observable characteristics, the obtained estimators can reduce the three possible sources of bias that arise from using non-experimental information: differences in the common support between the treatment and control groups; differences in the distribution of observable characteristics in both groups found within the area of common support; and differences in the results of interest owing to selection problems in non-observable characteristics.

The matching process for this study was conducted in three stages, following Imbens and Wooldridge (2009) and Imbens (2015). First, from the original ELP dataset we eliminated schools that were not found in the SIMAT dataset and hence lacked information on efficiency, cases with extreme variation in enrollment (reductions greater than 90% or increases above 70%), schools with a total enrollment less than 30 students, schools for which there are no available results on the Saber exams for 2009 (and which were not eligible for the program as a result), and schools that only received textbooks from ELP. Table A.4 summarizes this process and shows the proportion of observations eliminated in each trimming stage. In total, we eliminated 5% of schools in the treatment group (the clear majority were eliminated because of missing data for the Saber score in 2009) and 65% of the schools in the control group (nearly half of which either had less than 30 students or were not found in the SIMAT dataset in 2009 or 2014).

The second step was to identify the model that would maximize the match between the treatment and control groups. To do so, we first estimated a probability model of treatment for the complete sample of schools that were left after the trimming process described above. However, given that not every school in the sample had information on both efficiency and quality indicators, we estimated a different model for each type of outcome variable – efficiency and quality outcomes respectively. Table A.5 presents the results of the model used to estimate the propensity score for all schools using as covariates their observable characteristics in 2009. Consistent with the selection process, schools were more likely to enter the ELP program if they had a higher proportion of students with low scores on the Saber exams in 2009 (students who obtained an insufficient or minimum score on the Saber 9 exam), higher repetition and dropout rates in primary levels, higher proportion of vulnerable students (according to their socioeconomic

[13] Dropout was multiplied by $\frac{X_{rep}}{X_{des}}$ where X_{rep}= the maximum value for grade repetition within the corresponding bandwidth and X_{des}= the maximum for dropout within the corresponding bandwidth. In this way, the new, transformed variable for dropout has the same maximum and minimum values as grade repetition.

[14] It should be noted that, after creating the index and use it as running variable, the exercise becomes a partially fuzzy RDD (Battistin & Retore, 2008) and thus the requirement of monotonicity and changes the subpopulation are not required. We thank an anonymous referee for highlighting this point.



Graph 2. Distribution of the running variables close to their respective cut-off point (McCrary, 2008).
Notes: Each panel presents the McCrary (2008) test having as rnning variable rates of dropout (a), grade repetition (b), enrollment variation (c), and maximization index (d) in 2010, respectively. Source: MNE, authors' calculations.

stratum), and higher enrollment numbers as sought in the second selection phase. Schools were also more likely to participate if they were in CTEs marked as red or orange (critical) areas as defined by the MNE. Finally, following Crump., Hotz, Imbens, and Mitnik (2009), we trimmed the sample by eliminating schools with *pscore* values below 0.2 and above 0.8 ensuring a common support between treatment and control group. After this trimming we dropped 2682 treatment and 3346 control schools, ending up with a sample of 1299 benefited and 1360 control schools.[15]

Table A.6 shows the mean of the pre-treatment variables for this group of schools in the specified common support. As observed, the normalized differences of the observable characteristics used by the MNE in the selection process are statistically the same in 2009 between treatment and control groups as shown by the t-statistic and the value of the normalized difference, which is below 0.25, as recommended by Imbens and Wooldridge (2009). Furthermore, for most of all other observable characteristics of schools in 2009, except for the proportion of those that offer an afternoon class schedule, the differences between treatment and control schools are statistically equal to zero. Namely, in this range of common support, the treatment and control groups are comparable in both pre-treatment characteristics and in the pre-treatment outcomes of interest of test scores, dropout rates, and grade repetition.

With this sample of schools, we estimate a DID matching model that, as explained, allows us to control for time invariant observable and unobservable characteristics. For each dependent variable of interest, we estimate the average treatment on the treated on the first difference through specification (2) below:

$$\tau_{PMS\_DID} = \frac{1}{n_1}\sum_{i\in I_1\cap S_p}\left((Y^1_{i,2014} - Y^1_{i2009}) - \sum_{j\in I_0\cap S_p} W_{i,j}(Y^0_{j,2014} - Y^0_{j,2009})\right) \tag{2}$$

where I_1denotes the set of program participants, I_0 the set of non-program participants, n_1 the number of schools in the set I_1, S_p the common support and $W_{i,\ j}$ correspond to local weights under the Epanechnikov kernel. As with any DID matching estimation strategy, the identification assumption is that selection into the program is entirely explained by observable and unobservable characteristics that are constant in time. Given that we have the same information used in the selection process of benefited schools carried out by the MNE and McKinsey, this is a plausible assumption.

Finally, we also implement blocking regressions to estimate the treatment impacts of the ELP under a third alternative estimation methodology. As explained by Imbens (2015) blocking is similar to matching but allows a much more flexible weighting than the latter, while smoothing the propensity score within the subclasses and extreme values. We follow Imbens (2015) and divide the sample of schools within L intervals of the form [b_{l-1}, b_l) for l = 1, … .L where b_0 = 0.2 and b_l = 0.8. Imbens (2015) defines an indicator variable

[15] All regressions were also estimated dropping schools with pscore values below 0.1 and above 0.9. In this case, 2045 treatment and 2763 control schools are dropped ending up with a sample of 1882 benefited and 1997 control schools. The differences between pretreatment characteristics for most variables are also not statistically different from zero and the main results with this larger sample are quantitatively the same as those presented in this paper as can be observed in Table A.8.

$B_i(l) \in \{0, 1\}$ when the estimated propensity score for school i is in a range between b_{l-1} and b_l. Within each block, the average treatment impact is estimated using traditional linear regressions with pre-treatment control variables as shown in specification (2):

$$(\widehat{\alpha}_l, \widehat{\tau}_l, \widehat{\beta}_l) = argmin \sum_{i=1}^{N} B_i(l)^*\left((Y_{i,2014} - Y_{\ i,2009})_i - \alpha - \tau D_i - \beta^{'X_i}\right)^2 \tag{2}$$

These estimates give L estimators of τ (one for each block), which, as proposed by Imbens (2015), are averaged across the different groups in weighted fashion using the proportion of units in each block. To estimate the optimal number of L blocks we use the data-dependent methodology as proposed by Imbens (2015), which chooses L so that differences between the propensity score of treatment and control groups are sufficiently small and the number of observations in each group is sufficiently large.[16]

6. Results

This section summarizes the main results obtained using the RD design, the DID matching, and blocking regressions methods to estimate the impact of the ELP on measures of efficiency and quality of education in beneficiary schools.

6.1. Causal local and average impacts of the ELP

Table 3 presents the results from the RD estimations using the four alternative running variables that emerge from the first phase of the selection process of ELP schools. The dependent variables of interest at the school level are: dropout rate, grade repetition rate, and average math and language SABER 3 and 5 test scores for 2014. There are three columns for each running variable: (i) the first column displays the results from the second stage of RD estimations, which measure the local causal impact of the ELP on each variable of interest for schools around the cutoff point; (ii) the second column displays the results from the first stage of RD estimations, which establishes the probability of being treated if the school is eligible; and, (iii) the third column displays the number of schools that were included in the optimal bandwidth for each dependent variable.

For all the alternative running variables -except the 2009 repetition rate- the RD's first stage coefficient is statistically different from zero at a 1% significance. The discontinuous jump in the treatment probability is on average 75 percentage points when the index value is used, 40 percentage points for the 2010 dropout rate, and 30 for change in enrollment (2010–2009).

As for the impact of the treatment, the first column for each of these three running variables robustly shows that for all the outcome variables of interest, the causal local impact of the ELP for schools close to the cutoff point is statistically equal to zero. Results are maintained if we double the optimal bandwidth and use a larger number of schools for each alternative case – as presented in Table A.7 – suggesting that the lack of significance is not due to a low number of observations used in the estimations. Rather, results consistently suggest that the ELP has had no significant impact on measures of efficiency or quality in the schools included in the first phase of the selection process and are near the cutoff point for program eligibility. Three years after the start of its implementation in 2012, the ELP did not reduce dropout rates or grade repetition rates in treated schools; nor did it increase the standardized test scores of third or fifth grade students in the areas of mathematics and language that attend treated schools.

Table 4 presents the results obtained through the difference-in-difference (DID) matching and the blocking methodologies. In this case, the sample of schools used in the estimations comprises all beneficiary schools that lie within the range of the common support. Similar to the RD estimations, the DID matching model finds no significant impacts of the ELP on any of the changes between 2014 and 2009 quality indicators in institutions treated by the program. If anything, results suggest that the delta on grade repetition rates in these institutions increased by one percentage point. The last two columns of Table 4 confirm these results. Estimands obtained using blocking regressions suggest the ELP program had no impact on benefited schools except for a positive impact on dropout rates of the exact magnitude as that obtained using DID matching.[17]

Although results regarding an increase in repetition rates are unanticipated, such negative impacts of multilevel school programs have been previously found in other contexts. As shown in the literature section, such negative impacts were found for third grade students benefited by the program in China (Min, Yanqing, & Wenbin, 2012) and for urban students benefited by the program in Mexico (Paqueo & Lopez-Acevedo, 2003). Nonetheless, an explanation for such results could provide useful policy recommendations. Even though the stepwise matching estimation assured that treatment and control schools had the same observable characteristics in 2009, the last two rows in Table 4 go one step further and evaluate if there were any differences in changes in efficiency indicators between treatment and control schools during 2010 and 2009, before the ELP program was designed.[18] As observed, a positive and marginally significant increase in repetition rates (p-value 0.09) is found for these "would be" benefited schools. Even though we have no information that could help us explain this result, it could suggest that the ELP had no impact at all on repetition rates as it was unable to break such increase either and that there is not necessarily a negative impact on this efficiency rate.

6.2. Heterogeneous impacts of the ELP

The results presented so far systematically show that the ELP had no positive impact on quality indicators and, if anything, increased grade repetition rates. Yet, it is possible that the program has been effective in certain types of schools, depending on their structure or on the intrinsic characteristics of the program. Of the three main program components (materials, teacher mentoring, and principal training), the most important in terms of design and resources invested was the second one. Through teacher mentoring, the program sought to help teachers establish effective teaching and classroom practices to improve the use of classroom time and enhance students' learning. Although tutors were selected through a merit-based process and were trained for their tasks, it is possible that specific characteristics of these trainers differentially affected the results. Similarly, it is possible that school characteristics could also affect the program's implementation and thereby affect the results. For example, it is likely that tutor visits to rural schools were more difficult and infrequent than for urban schools. Alternatively, a tutor's potential impact may be greater in smaller schools where they can concentrate their efforts on a smaller number of teachers and students.

To evaluate these possibilities, we estimated heterogeneous impacts through the DID matching models, according to specific characteristics of schools and tutors. For tutors, we analyzed whether their focus area – mathematics or language – had a differential impact on the measures of efficiency and quality.[19] For schools, we evaluated whether institutions

[16] Matching estimations were undertaken using the psmatch2 Stata command. For blocking regressions, we used pstrata and mmws Stata commands.

[17] Table A.8 presents these same results using all schools with a pscore between 0.1 and 0.9 which increases total number of observations used in estimations by almost 45%. All results remain the same. Table A.9 presents the number of optimal blocks obtained for each sample and summarizes their main descriptive statistics regarding balance of the pscore within each of them.

[18] Such an exercise for Saber 5 and Saber 3 exams is not possible as after 2009 Saber 5 was only applied again in 2012 and Saber 3 was first applied in 2012.

[19] Most tutors had experience in one of these two areas of knowledge, as the selection process explicitly sought teachers with this background.

Table 3

Causal impact of the ELP on schools selected in Stage 1 obtained through a regression discontinuity design.

	Running variable											
	Max index			Dropout rate (2010)			Repetition rate (2010)			Change in enrollment (2010–2009)		
Outcome of interest	Beta Second Stage (1)	Beta First Stage (2)	No. of observations (3)	Beta Second Stage (4)	Beta First Stage (5)	No. of observations (6)	Beta Second Stage (7)	Beta First Stage (8)	No. of observations (9)	Beta Second Stage (7)	Beta First Stage (8)	No. of observations (9)
Primary Dropout (2014)	−0.008	.807***	174	0.027	.345**	325	−0.281**	.131	497	0.002	−0.283***	1138
	(0.045)	(0.115)		(0.047)	(0.154)		(0.142)	(0.087)		(0.030)	(0.054)	
Primary Grade Repetition (2014)	0.025	.808***	171	0.011	.398***	428	0.092	.14*	522	−0.014	−0.278***	1191
	(0.049)	(0.115)		(0.047)	(0.127)		(0.126)	(0.083)		(0.039)	(0.052)	
Math Saber 5 test score (2014)	−14.573	.784***	196	−31.192	.377***	324	50.971	.078	483	−0.996	−0.236***	1077
	(27.724)	(0.105)		(30.681)	(0.142)		(123.021)	(0.083)		(29.486)	(0.053)	
Language Saber 5 test score (2014)	−35.060	.786***	184	−37.977	.351**	282	44.026	.083	315	15.168	−0.251***	351
	(27.006)	(0.111)		(25.558)	(0.154)		(91.415)	(0.082)		(20.501)	(0.055)	
Math Saber 3 test score (2014)	−12.494	.726***	261	−25.137	.475***	491	−27.070	.11	465	35.116*	−0.277***	1189
	(35.819)	(0.085)		(20.448)	(0.108)		(92.631)	(0.086)		(21.290)	(0.047)	
Language Saber 3 test score (2014)	−43.694	.739***	256	−38.356	.356**	298	−52.217	.109	466	18.159	−0.269***	1195
	(35.253)	(0.087)		(33.390)	(0.145)		(76.444)	(0.087)		(22.180)	(0.053)	

Note: The table presents the results of the RD estimations using the four alternative running variables for schools selected for the ELP program in the first stage. For each RD model, the first column presents the impact of the ELP program on the dependent variables of interest (beta second stage), the second column presents the change in treatment probability (first stage) and the third column presents the number of observations included in the optimal bandwidth. All estimations carried out using the "rdrobust" package of Cattaneo et al. (2014). Robust standard errors presented in parenthesis. *** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.

Table 4
Average causal impact of ELP on benefited schools under DID matching and PSM blocking estimations.

Outcome of interest	Results under DID matching		Results under PSM Blocking	
	Beta DID matching (1)	No. of observations (2)	Beta PSM Blocking (3)	No. of observations (4)
Delta primary dropout rate (2014–2009)	0.003	2659	0.004	2659
	(0.002)		(0.003)	
Delta primary grade repetition rate (2014–2009)	0.010***	2659	0.013**	2659
	(0.003)		(0.004)	
Delta math Saber 5 test score (2014–2009)	−0.79	2386	−0.49	2386
	(1.79)		(2.28)	
Delta language Saber 5 test score (2014–2009)	−2.45*	2386	−2.78	2386
	(1.45)		(1.91)	
Math Saber 3 test score (2014)	−1.45	2386	−2.30	2239
	(1.65)		(1.77)	
Language Saber 3 test score (2014)	−1.31	2386	−2.58	2230
	(1.59)		(1.75)	
Delta primary dropout rate (2010–2009)	0.004*	2659	0.006*	2659
	(0.002)		(0.003)	
Delta primary grade repetition rate (2010–2009)	0.005*	2659	0.009**	2659
	(0.003)		(0.004)	

Note: Columns (1)-(2) present the main DID matching results using the psmatch2 package. Columns (3)-(4) present the main PSM blocking results using the pstrata and mmws command. Controls in these two methodologies include pre-treatment characteristics of 2009 such as dropout rate, repetition percentage of students with an insufficient Saber 5 score, number of establishments of the school, area -urban-rural-, and ETC category. Bootstrap standard errors in parenthesis *** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.

Table 5
DID matching heterogeneous impacts of the ELP, according to school tutor characteristics and program intensity.

Outcome of interest	Schools' characteristics				Tutor's characteristics					
	Location		Number of establishments		Area of expertise		Number of visits		Number of days in school	
	Urban	Rural	> median	< = median	Language	Math	> median	< = median	> median	< = median
Primary dropout	0.005	−0.001	−0.005	0.005	−0.002	−0.003	−0.002	0.001	−0.003	0.002
	(0.006)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Primary grade repetition	0.019**	0.009**	0.004	0.015***	0.014***	0.007*	0.007*	0.013***	0.006	0.013***
	(0.008)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Math test score (5th grade)	−1.85	0.71	2.59	−4.46*	−3.78*	−0.41	−1.04	−2.18	0.28	−3.72
	(3.040)	(2.700)	(3.020)	(2.470)	(2.270)	(2.410)	(2.420)	(2.680)	(2.200)	(2.540)
Language test score (5th grade)	−3.78	−0.35	1.77	−5.35**	−5.58***	−1.05	−1.62	−4.10*	−1.48	−4.20*
	(2.610)	(2.190)	(2.630)	(2.410)	(1.920)	(2.620)	(1.730)	(2.270)	(1.960)	(2.190)
Math test score (3rd grade)	1.82	−1.59	−0.57	−1.52	−2.87	−3.22	−1.62	−2.58	−0.6	−4.07*
	(2.670)	(2.410)	(2.800)	(2.220)	(2.390)	(2.590)	(2.190)	(2.040)	(2.000)	(2.230)
Language test score (3rd grade)	4.64*	−3.64	−1.52	−1.28	−2.76	−2.92	−0.1	−3.45*	−0.34	−3.44
	(2.510)	(2.510)	(2.950)	(2.560)	(2.600)	(2.780)	(2.060)	(2.030)	(2.100)	(2.570)

Note: Columns present the main DID matching heterogeneity results using the psmatch2 package. Controls include Controls in these two methodologies include pre-treatment characteristics of 2009 such as dropout rate, repetition percentage of students with an insufficient Saber 5 score, number of establishments of the school, area -urban-rural-, and ETC category. Robust standard errors in parenthesis. *** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.

located in urban versus rural areas, or those with a higher number of locations[20], might show differential impacts.

To estimate these heterogeneous impacts, we divided the treatment and control groups into different sub-samples. Specifically, for the categorical variables (whether the school is in an urban or rural area; the tutor's academic area) we divided the sample into the characteristics to be analyzed. When estimating heterogeneous impacts with continuous variables (number of school locations), we calculated the median within the treatment group and ran regressions for the schools above and below the median. We estimated a new propensity score separately for each subsample following the same procedures as with the full sample.[21]

Table 5 summarizes the results obtained for all treated schools (that fall into the common support) in each subsample under the DID matching estimations. Results are consistent with those from Tables 3 and 4. We are unable to reject the null hypothesis that the impact of the ELP on education outcomes is zero. In fact, results of an increase of grade repetition rates for almost all subsamples is maintained and a negative impact on Saber 5 language score seems to emerge. Rural small schools

[20] In the Colombian education system public schools may have one or more locations. Many times, for schools with multiple locations, each location offers a given schooling level (i.e., primary or high school). However, it is possible that more than one location offers a given level of education. If many offer primary education, the ELP tutors may need to divide their time among several locations, reducing the impact of the program in each specific site.

[21] Of course, by dividing the sample and estimating the DID matching models on these sub-samples, we need to confirm that balance is still attained and that the new propensity score estimated for each subsample is thus appropriate. Although not presented in this paper, we estimated a new matching procedure for each subsample and these criteria are maintained, assuring that a good match was effectively attained. Details on these ten matching estimations are all available upon request from the authors.

Table 6
The ELP in the field – implementation details.
Source: ELP program field data. Authors' calculations.

	Mean or %	Standard deviation	Minimum	Maximum	Number of schools
Number of tutor visits per year					
2012	4.73	2.46	0	7	131
2013	6.22	1.39	0	7	190
2014	5.79	1.72	0	7	188
Number of personalized tutor meetings per year					
2013	3.01	2.50	0	7	169
2014	2.56	2.36	0	7	172
Activities conducted by tutors during the mentorship process (%)					
Tutor-assisted lesson planning	0.71	0.45	0	1	197
Class observations	0.58	0.49	0	1	197
Reflection and feedback on teaching practices	0.78	0.41	0	1	197
Theoretical classes	0.86	0.34	0	1	197
Model classes	0.54	0.50	0	1	197
Reviewing results from diagnostic exams	0.85	0.35	0	1	198
Reviewing results from SABER exams	0.81	0.39	0	1	199

that had a tutor whose area of expertise was language saw a decrease on average language scores between 5 and 6 points, corresponding to a reduction of 2% of the mean in 2010 – before ELP implementation.

7. Implementation deficiencies as possible explanations for results

The analysis of the results obtained in the RD, DID matching, and blocking estimations suggests that the ELP program, by 2014, had not accomplished its short-run objectives. Understanding the reasons for these results, which could be explained by a variety of channels, is crucial for policymakers. In this section, we evaluate whether deficiencies or heterogeneity in the program's implementation may be part of the explanation. To do so, we turn to data we collected in the field in 2014 of a representative sample of 200 treated and 200 control schools.

Perhaps the aspect that best explains the failure to find significant impacts of the ELP on classroom practices, and consequently on measures of quality and efficiency, is related to the implementation of the most relevant component of the program: the personalized visits that tutors made to schools and the activities that tutors conducted with teachers. Two important details emerge from an analysis of the information collected in the field.

The first notable fact is the low number of visits that treated schools received from tutors and the low level of individualized interaction that tutors had with teachers. In its initial design, it was expected that tutors would provide in-school training to teachers during four group sessions and four individualized sessions per month. That is, tutors would visit schools and interact with the teachers at least eight times per month during the academic year either individually or in groups, with the expectation that this would enable tutors to impact educators' teaching practices and improve the quality of education provided in these schools. Once in the field however, the actual number of visits and individualized attention was significantly lower than originally planned. According to Table 6 the total average number of tutor visits to ELP schools per year was 4.7 in 2012, 6.2 in 2013, and 5.8 by August of 2014. In fact, the maximum number of times that a tutor visited a given school was seven, which is less than the total visits expected in two months.

It is also worrisome that teachers reported having individualized sessions with the tutor in less than half of these visits. The mean number of individual sessions for 2013 was 3.0 for the entire year, and 2.6 by August of 2014, substantially below the goal of four monthly sessions. It is unlikely that classroom teaching practices, and hence the quality of education received by students, could be effectively modified in 3 annual sessions, a figure that is far below what was specified in the original program's design.

Not surprisingly, results from García et al. (2017) using this same dataset find that the ELP had only a minor impact on teachers' classroom practices. That analysis found positive impacts for ELP teachers on classroom planning time in language and mathematics, greater use of digital tools to register lesson plans, and greater (self-reported) knowledge of language and mathematics standards than teachers in non-ELP schools. It found no differences, however, in the use of student-centered teaching practices, the use of evaluations to provide feedback to students or parents, and the use of strategies to establish either clear classroom norms or to improve student participation. This is again consistent with the results in the last four columns of Table 5, which evaluate heterogeneous impacts according to the number of visits and the days in school by the tutor. Analyzing the impact of the tutors' visits according to whether the number of visits made by each school's tutor was above or below the median does not change the main results.

The second aspect that must be highlighted, which could also explain the lack of impact is the heterogeneity in program implementation, an aspect that could be related to the absence of standardized protocols in many core activities such as teacher training and principal support. Even in the small number of visits that each tutor made to each ELP school, activities conducted by them with teachers varied significantly across schools. Originally, the design of the program envisioned training based on "hands-on activities" and development of teaching practices, with less focus on lectures. However, as shown in Table 6, the most frequent activities were theoretical classes (86%), followed by the review of results from diagnostic exams (85%) and reviewing results or previous exams of Saber standardized tests (81%). Contrary to what was expected, the least frequent activities that tutors undertook with teachers were model classes (54%), class observations (58%), and collaborative lesson planning (71%).

Given the low number of visits and personalized sessions, as well as the lack of guidelines for teacher training, it is not surprising that the program did not change the classroom practices of teachers, the program's central objective from the outset. With the lack of evidence for any impact on classroom practices, it is not surprising that the measures of school quality and efficiency also experienced no impact.

8. Conclusions

This paper presents the results of the impact evaluation of Colombia's Everyone Learning Program on dropout rates, grade repetition rates, and standardized test scores in mathematics and language of students in benefited schools. The quasi-experimental estimations we conduct with census data allow us to conclude that, at least in the short run, the ELP program has not achieved any impact on the main proposed objective of the program: improving the quality of education offered by Colombia's most vulnerable schools.

The results are robust to different types of estimations (regression discontinuity design, difference-in-difference matching, and blocking) and to the use of different groups of schools. An analysis of heterogeneous impacts suggests that the absence of significant results is common across all alternative groups independent of both school and tutor characteristics. Data collected from the field suggests that the explanation may be based on deficiencies in program implementation, which was not carried out as initially envisioned in the program design. The intensity of teacher mentoring sessions was minimal, and the sessions that were conducted were not carried out according to a standard methodology for all schools, nor were they based on standards for content and pedagogical goals. This suggests the need to reinforce pedagogical training in teacher mentoring sessions in such a way as to directly impact the classroom environment.

Results suggest that even though multilevel school reforms have proven to bring positive impacts to benefited schools in both developing and developed countries, the achievement of positive results is not automatic. The Colombian experience shows that differences between original design and actual implementation can be critical for the impact of a program. Ultimately, for the ELP, such deficiencies meant that no significant changes in teaching practices were achieved, and therefore the program saw no improvement in quality or efficiency education indicators.

Acknowledgments

We gratefully acknowledge all the information provided by the Ministry of Education and the ELP team, particularly Lina María López, Juan Carlos Hernández, Paola Trujillo, Manuel Rivera, and Rodrigo Sepúlveda. Special thanks to Xiomara Pulido who provided valuable research assistantship in the last stages of the paper. Finally, we thank the Economics Department and CEDE at Universidad de los Andes for the financial support provided for the elaboration of this paper. Finally, we thank the thorough review and comments of the Editor and two anonymous referees. All remaining errors are ours.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2018.07.001.

Appendix

Table A.1
Pre-treatment summary statistics for schools selected in each phase.
Source: MNE, authors' calculations.

	(Schools originally selected by MNE)			(Schools selected by McKinsey)			(Schools selected by CTE)		
	Mean	Standard deviation	N	Mean	Standard deviation	N	Mean	Standard deviation	N
Enrollment variation 2009–2010 (Assignment variable)	−0.0350	0.136	1367	0.00733	0.104	987	0.0470	0.591	1847
Dropout (Assignment variable)	0.0936	0.0563	1367	0.0500	0.0414	883	0.0694	0.0584	1756
Grade repetition (Assignment variable)	0.0685	0.0446	1367	0.0454	0.0283	883	0.0548	0.0657	1756
Math average Saber5 2009	264.5	38.48	1281	275.6	30.07	986	282.3	42.35	1643
Language average Saber5 2009	265.5	32.41	1268	277.2	26.08	987	279.1	33.58	1634
Vulnerable population	0.682	0.198	1366	0.667	0.171	984	0.695	0.215	1841
Rural schools	0.621	0.485	1367	0.260	0.439	988	0.733	0.442	1900
Number of teachers	22.41	18.77	1350	39.98	22.87	889	17.37	16.27	1798
Enrollment 2010 R166	967.7	897.5	1308	1880	1061	869	718.7	755	1739
Full-day schedule	0.141	0.348	1366	0.146	0.353	988	0.342	0.474	1850
Morning schedule	0.967	0.179	1366	0.980	0.141	988	0.842	0.365	1850
Afternoon schedule	0.592	0.492	1366	0.849	0.358	988	0.431	0.495	1850
Evening schedule	0.442	0.497	1366	0.613	0.487	988	0.390	0.488	1850
Weekend schedule	0.595	0.491	1366	0.486	0.500	988	0.564	0.496	1850

Table A.2
Number of observations eliminated in the trimming process for the Regression Discontinuity estimation (RD).
Source: MNE, authors' calculations.

	Treatment		Control		Total	
	number	%	number	%	number	%
Initial total	4,255	100.0%	13,160	100.0%	17,415	100.0%
Trimming steps (eliminated observations)						
Schools selected in Step 2 and Step 3	2888	67.9%	409	3.1%	3297	18.9%
Schools not found in R166	0	0.0%	2919	22.2%	2919	16.8%
Schools with no baseline information	0	0.0%	1001	7.6%	1001	5.7%
Schools that only received books	0	0.0%	1048	8.0%	1048	6.0%
Schools with high Saber 5 scores	0	0.0%	1436	10.9%	1436	8.2%
Schools in CTE where no school in Step 1 were located	0	0.0%	5091	38.7%	5091	29.2%
Total schools eliminated before RD	2,888	67.9%	11,904	64.7%	14,792	50.1%
Total schools used for the RD	1,367	32.1%	1256	35.0%	2623	50.0%

Table A.3
Continuity of pre-treatment schools' characteristics around the cutoffs points for all four running variables.

	Running variable							
	Max		Dropout		Repetition		Changes in enrollment	
Variables	Beta second stage	Observations	Beta second stage	Observations	Beta second stage	Observations	Beta second stage	Observations
Dropout rate 2010 (R166)	0.021	318	0.042	263	0.006	912	0.021	1540
	(0.050)		(0.045)		(0.022)		(0.016)	
Grade repetition 2010 (R166)	0.031	203	0.007	334	0.006	646	0.016	1303
	(0.020)		(0.021)		(0.012)		(0.012)	
Math average Saber 5 score (2009)	−1.152	298	8.126	372	−1.784	632	9.511*	1171
	(17.801)		(10.149)		(9.682)		(5.765)	
Language average Saber 5 score (2009)	6.057	287	12.957	277	4.459	636	1.300	1329
	(15.916)		(9.828)		(7.249)		(4.873)	
Percentage of teachers with high school degree	−10.017	433	1.409	276	−1.264	800	4.465*	1468
	(7.098)		(5.524)		(3.538)		(2.445)	
Percentage of teachers with normal education	−1.835	303	1.567	344	4.156	688	3.913	1568
	(12.060)		(6.550)		(3.529)		(2.391)	
Percentage of teachers with undergraduate degree	−3.166	330	−11.046	319	−2.904	1111	−0.407	1646
	(10.871)		(9.715)		(4.491)		(3.640)	
Percentages of teachers with 2-year degrees	2.209	292	7.631*	501	−0.850	721	−2.852*	1623
	(3.979)		(4.242)		(3.032)		(1.456)	
Percentage of teachers with post-graduate degree	7.491	470	1.101	357	−1.231	951	−4.174	1581
	(6.076)		(6.615)		(3.816)		(2.638)	
Enrollment 2010 (R166)	542.039***	372	90.825	356	23.980	729	−278.570***	1465
	(170.122)		(200.208)		(165.949)		(91.642)	
Total teachers	15.410***	315	5.234	491	3.895	971	−5.330***	1869
	(5.261)		(4.506)		(3.320)		(1.706)	
Full-day Schedule	0.106	447	0.047	527	0.085	641	0.003	1559
	(0.103)		(0.095)		(0.072)		(0.038)	
Morning Schedule	0.032	387	0.079	457	−0.039	669	−0.013	1922
	(0.097)		(0.081)		(0.035)		(0.025)	
Afternoon Schedule	0.272	309	0.083	487	0.047	726	−0.084*	1885
	(0.179)		(0.129)		(0.099)		(0.048)	
Night Schedule	0.340***	306	0.155	305	0.127	808	−0.099**	1798
	(0.121)		(0.149)		(0.091)		(0.048)	
Number of school locations (MNE)	0.981	347	−0.032	475	0.312	810	−2.841***	1520
	(1.595)		(1.329)		(0.786)		(0.639)	
Unsatisfied Basic Needs Index (MNE)	−9.912	342	−1.460	348	−3.278	780	4.899**	1723
	(8.143)		(7.237)		(4.158)		(2.368)	

Note: The table presents the results of the RD estimations to test continuity in pre-treatment characteristics of treated and control schools before the ELP program was implemented using the four alternative running variables. For each one, the first column presents the coefficient testing differences between the two groups of schools and the third column presents the number of observations included in the optimal bandwidth. All estimations carried out using the "rdrobust" package of Cattaneo et al. (2014). Robust standard errors presented in parenthesis. *** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.

Table A.4
Number of observations eliminated in the trimming process for the propensity score estimation used in the Propensity Score Matching (PSM).
Source: MNE administrative data, authors' calculations.

	Treatment		Control		Total	
	number	%	number	%	number	%
Initial Total	4,255	100%	13,160	100%	17,415	100%
Trimming steps (eliminated observations)						
Schools not found in SIMAT	0	0%	2,919	22.18%	2,919	16.76%
Extreme variation in school enrollment (< −90% ó > 0.7)	31	0.73%	272	2.07%	303	1.74%
2010 school enrollment less than 30	40	0.94%	2746	20.87%	2786	16.00%
Schools with incomplete information for variables of result 2009	142	3.34%	1529	11.62%	1671	9.60%
Schools that only received books	0	0%	1,049	7.97%	1,049	6.02%
Total eliminated before PSM	213	5.01%	8,515	64.70%	8,728	50.12%
Total used for the PSM	4,042	95%	4,645	35%	8,687	50%

Table A.5
Estimated probability of ELP participation after initial trimming.

	Probit for the sample having dropout and repetition as outcome	Probit for the sample having Saber scores as outcome
Score SABER Language 2009	0.004***	0.003
	(0.002)	(0.002)
Score SABER Mathematics 2009	−0.004**	−0.004**
	(0.002)	(0.002)
Inferior Level in Language (% of students)	0.007*	0.006
	(0.004)	(0.004)
Inferior Level in Mathematics (% of students)	0.019***	0.019***
	(0.003)	(0.003)
Dropout (school total) 2009	−1.630***	−1.473***
	(0.402)	(0.455)
Dropout primary 2009	1.376***	1.302***
	(0.450)	(0.496)
Grade repetition (school total) 2009	−0.972*	−1.144*
	(0.585)	(0.640)
Grade repetition primary 2009	1.056*	0.979
	(0.569)	(0.617)
Rural area school	−0.042	−0.038
	(0.115)	(0.119)
Rural population - 2010	−0.001	−0.001
	(0.002)	(0.002)
Strata (vs. Strata 3)		
Strata 1	0.470***	0.443***
	(0.099)	(0.105)
Strata 2	0.556***	0.531***
	(0.071)	(0.074)
Number of school locations	0.070***	0.074***
	(0.011)	(0.012)
Number of school locations ˆ2	−0.002***	−0.002***
	(0.000)	(0.001)
Enrollment 2009	0.000	0.000
	(0.000)	(0.000)
Enrollment primary 2009	0.000	0.000
	(0.000)	(0.000)
Enrollment (vs quintile 5)		
Quintile 1	−2.101***	−2.346***
	(0.175)	(0.215)
Quintile 2	−1.334***	−1.383***
	(0.136)	(0.142)
Quintile 3	−0.674***	−0.733***
	(0.120)	(0.124)
Quintile 4	−0.212**	−0.262***
	(0.092)	(0.096)
Morning schedule	−0.095	−0.058
	(0.061)	(0.064)
Afternoon schedule	0.371***	0.398***
	(0.074)	(0.077)
Evening schedule	−0.138***	−0.127**
	(0.052)	(0.055)
Red (critical) status	1.063***	1.070***
	(0.062)	(0.067)
Green (acceptable) status	−1.262***	−1.260***
	(0.067)	(0.070)
Observations	6,259	5,584

Note: The table presents the results of a propensity score estimation, having as control variables the RD estimations to test continuity in pre-treatment characteristics of treated and control schools before the ELP program was implemented using the four alternative running variables. For each one, the first column presents the coefficient testing differences between the two groups of schools and the third column presents the number of observations included in the optimal bandwidth. All estimations carried out using the "rdrobust" package of Cattaneo et al. (2014). Robust standard errors presented in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Source: MNE, authors' calculations.

Table A.6
Pre-treatment characteristics of Treatment (ELP) and Control (non-ELP) group for the DID matching estimations.
Source: MNE administrative data, authors' calculations.

	Dropout and repetition sample				Saber scores sample			
	Mean		*t*-test	Normalized difference	Mean		*t*-test	Normalized difference
	Treatment	Control			Treatment	Control		
pscore	0.579	0.578	0.001 (0.006)	0.00564	0.578	0.577	0.001 (0.007)	0.00679
Score SABER language 2009	289.7	289.7	−0.012 (1.180)	−0.000396	290.7	289.6	1.051 (1.208)	0.0351
Score SABER mathematics 2009	292.1	292.9	−0.704 (1.385)	−0.0195	293.1	292.7	0.426 (1.416)	0.0121
Inferior level in language (% of students)	19	19.13	−0.130 (0.452)	−0.0110	18.57	18.98	−0.405 (0.457)	−0.0358
Inferior level in mathematics (% of students)	38.26	38.34	−0.083 (0.717)	−0.00445	37.72	38.04	−0.316 (0.728)	−0.0175
Dropout total 2009	0.134	0.137	−0.003 (0.003)	−0.0414	0.132	0.136	−0.004 (0.003)	−0.0444
Dropout primary 2009	0.0740	0.0777	−0.004 (0.003)	−0.0502	0.0734	0.0773	−0.004 (0.003)	−0.0532
Grade repetition 2009	0.0983	0.101	−0.003 (0.003)	−0.0327	0.0981	0.0989	−0.001 (0.003)	−0.0103
Grade repetition primary 2009	0.0962	0.0996	−0.003 (0.004)	−0.0367	0.0962	0.0971	−0.001 (0.004)	−0.00979
Rural area school	0.627	0.618	0.009 (0.019)	0.0190	0.626	0.615	0.011 (0.020)	0.0222
Rural population - 2010	62.70	61.54	1.163 (1.598)	0.0279	62.79	61.47	1.321 (1.680)	0.0317
Strata (vs. Strata 3)								
Strata 1	0.587	0.590	−0.003 (0.019)	−0.00562	0.588	0.586	0.002 (0.020)	0.00448
Strata 2	0.183	0.209	−0.026* (0.015)	−0.0663	0.182	0.212	−0.030* (0.016)	−0.0746
Number of school locations	4.665	4.686	−0.021 (0.151)	−0.00530	4.845	4.911	−0.066 (0.167)	−0.0159
Number of school locations ˆ2	37.01	37.50	−0.493 (2.391)	−0.00792	41.44	40.50	0.937 (3.478)	0.0109
Enrollment 2009	931.7	876.5	55.246 (34.227)	0.0619	939.5	895	44.483 (36.820)	0.0487
Enrollment primary 2009	403.4	384.7	18.732 (13.528)	0.0531	408.1	394.6	13.469 (14.757)	0.0368
Enrollment (vs. quintile 5)								
Quintile 1	0.0154	0.0222	−0.007 (0.005)	−0.0497	0.0114	0.0163	−0.005 (0.005)	−0.0416
Quintile 2	0.259	0.269	−0.010 (0.017)	−0.0237	0.251	0.270	−0.019 (0.018)	−0.0444
Quintile 3	0.338	0.335	0.004 (0.018)	0.00780	0.344	0.344	−0.000 (0.019)	−6.88e-05
Quintile 4	0.201	0.212	−0.010 (0.016)	−0.0254	0.208	0.204	0.004 (0.016)	0.0100
Morning schedule	0.374	0.388	−0.014 (0.019)	−0.0291	0.386	0.391	−0.005 (0.020)	−0.0107
Afternoon schedule	0.817	0.812	0.005 (0.015)	0.0138	0.805	0.798	0.007 (0.016)	0.0174
Evening schedule	0.513	0.510	0.004 (0.019)	0.00701	0.508	0.522	−0.014 (0.020)	−0.0273
Red (critical) status	0.129	0.132	−0.004 (0.013)	−0.0111	0.120	0.119	0.000 (0.013)	0.00115
Green (acceptable) status	0.109	0.102	0.006 (0.012)	0.0208	0.121	0.113	0.008 (0.013)	0.0246
Observations	1360	1299	2,718	1228	1158	2,458		

Note: * $p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table A.7

Causal impact of the ELP on schools selected in Stage 1 obtained through a Regression Discontinuity Design (double optimal bandwith).

	Running variable											
	Max Index			Dropout rate			Repetition rate			Change in enrollment		
Outcome of interest	Beta Second Stage	Beta First Stage	No. of observations	Beta Second Stage	Beta First Stage	No. of observations	Beta Second Stage	Beta First Stage	No. of observations	Beta Second Stage	Beta First Stage	No. of observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Primary dropout	−0.002	.760***	359	0.017	.37***	627	−0.190**	.14*	827	−0.002	−0.28***	1391
	(0.045)	(0.09)		(0.035)	(0.13)		(0.074)	(0.08)		(0.028)	(0.05)	
Primary grade repetition	0.033	.770***	353	0.012	.44***	812	0.067	.15**	873	−0.009	−0.28***	1407
	(0.049)	(0.1)		(0.039)	(0.11)		(0.099)	(0.07)		(0.036)	(0.05)	
Math test score (5th grade)	−5.249	.740***	387	−25.523	.42***	613	76.384	.1	812	8.526	−0.24***	1372
	(25.446)	(0.09)		(25.382)	(0.12)		(79.613)	(0.07)		(26.713)	(0.05)	
Language test score (5th grade)	−28.649	.740***	362	−26.384	.38***	549	18.416	.11	524	22.563	−0.25***	451
	(21.127)	(0.09)		(17.823)	(0.13)		(39.167)	(0.07)		(16.427)	(0.05)	
Math test score (3rd grade)	−3.231	.700***	513	−25.943	.47***	912	−16.519	.11	793	35.498*	−0.28***	1370
	(31.494)	(0.07)		(21.154)	(0.11)		(71.282)	(0.08)		(20.520)	(0.05)	
Language test score (3rd grade)	−34.089	.700***	498	−38.705	.4***	573	−31.565	.11	783	18.202	−0.28***	1381
	(32.442)	(0.07)		(26.259)	(0.13)		(59.406)	(0.08)		(21.002)	(0.05)	

Note: The table presents the results of the RD estimations using the four alternative running variables for schools selected into the ELP program in the first stage. For each RD model, the first column presents the impact of the ELP program on the dependent variables of interest (beta second stage), the second column presents the change in treatment probability (first stage) and the third column presents the number of observations included in the optimal bandwidth. All estimations carried out using the “rdrobust” package of Cattaneo et al. (2014). Robust standard errors presented in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Source: MNE, authors' calculations.

Table A.8
Average causal impact of the ELP on benefited schools under DID matching and PSM blocking estimations (0.1 < Propensity score < 0.9).

Outcome of interest	Results under DID matching		Results under PSM blocking	
	Beta DID matching (1)	No. of observations (2)	Beta PSM blocking (3)	No. of observations (4)
Delta primary dropout rate (2014–2009)	0.001 (0.003)	3868	−0.001 (0.003)	3879
Delta primary grade repetition rate (2014–2009)	0.013*** −0.003	3868	0.011*** −0.004	3879
Delta math Saber 5 test score (2014–2009)	−1.19 (1.650)	3462	0.59 (2.033)	2905
Delta language Saber 5 test score (2014–2009)	−2.61* (1.490)	3462	−1.982 (1.740)	2905
Math Saber 3 test score (2014)	−1.66 (1.730)	3462	−1.292 (1.680)	2721
Language Saber 3 test score (2014)	−1.02 (1.750)	3462	−1.786 (1.701)	2710
Delta primary dropout rate (2010–2009)	−0.003 (0.002)	3,868	−0.003 (0.003)	3879
Delta primary grade repetition rate (2010–2009)	0.010*** (0.003)	3868	0.007* (0.004)	3879

Note: Columns (1)-(2) present the main DID matching results using the psmatch2 package. Columns (3)-(4) present the main PSM blocking results using the pstrata and mmws command. Controls in these two methodologies include pre-treatment characteristics of 2009 such as dropout rate, repetition percentage of students with an insufficient Saber 5 score, number of establishments of the school, area -urban-rural-, and ETC category. Bootstrap standard errors in parenthesis *** p < 0.01, ** p < 0.05, * p < 0.1. Source: MNE, authors' calculations.

Table A.9
Propensity score by blocks.

Subclasses	*p*-score		Number of control units	Number of treated units	Average *p*-score		Normalized difference
Dropout and repetition sample							
	Min.	Max.			Control	Treated	
1	0.100	0.161	306	47	0.131	0.135	0.216
2	0.162	0.228	294	59	0.192	0.196	0.201
3	0.229	0.313	251	101	0.267	0.273	0.224
4	0.313	0.405	222	131	0.358	0.363	0.191
5	0.406	0.491	210	143	0.447	0.449	0.0783
6	0.491	0.573	174	178	0.533	0.534	0.0540
7	0.574	0.650	145	208	0.611	0.613	0.0773
8	0.650	0.730	104	249	0.690	0.691	0.0333
9	0.730	0.793	74	278	0.762	0.762	−0.0193
10	0.793	0.852	67	286	0.820	0.822	0.125
11	0.852	0.900	35	317	0.877	0.877	−0.0307
Saber score sample							
1	0.150	0.210	247	44	0.178	0.176	−0.103
2	0.210	0.285	219	71	0.243	0.245	0.126
3	0.285	0.369	190	101	0.324	0.326	0.0781
4	0.369	0.446	169	121	0.411	0.406	−0.226
5	0.446	0.528	159	132	0.484	0.489	0.192
6	0.528	0.601	130	160	0.563	0.567	0.163
7	0.601	0.671	110	181	0.633	0.637	0.191
8	0.671	0.740	80	210	0.707	0.707	−0.005
9	0.740	0.797	62	229	0.765	0.768	0.175
10	0.797	0.85	53	237	0.822	0.823	0.0237

References

Barrera, F., Maldonado, D., & Rodríguez, C. (2014). Calidad de la Educación Básica y Media en Colombia: Diagnóstico y Propuestas. In A. Montenegro, & M. Meléndez (Eds.). *Equidad y movilidad social* (pp. 239–328). Bogotá: Universidad de los Andes.

Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics, 142*(2), 715–730.

Bellei, C. (2011). *Using a regression-discontinuity design to evaluate the causal impact of a compensatory educational program: evidence on the effect of in-School professional development in low-performing schools in chile.* Santiago de, Chile: Center for Advanced Research in Education, University of Chile.

Cattaneo, M. D., Jansson, M., & Ma, X. (2016a). *Manipulation testing based on density discontinuity.* University of Michigan Working Paper.

Cattaneo, M. D., Jansson, M., & Ma, X. (2016b). *Simple local regression distribution estimators with an application to manipulation testing.* University of Michigan Working Paper.

Cattaneo, M. D., Calonico, S., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica, 82*(6), 2295–2326.

Cerdan-Infantes, P., & Vermeersch, C. (2007). *Policy Research Working Paper Series No 4167*Washington DC: The World Bank.

Chay, K., McEwan, P., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *The American Economic Review, 95*(4), 1237–1258.

Crump., R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*(1), 187–199.

García, S., Harker, A., Figueroa, M., Gómez-Echeverry, S., & Rojas, M. P. (2017). *The dire* *role of program design and fidelity of Implementation: lessons from a large-scale On-site teacher training program,.* Mimeo.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2013). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In Paul Glewwe (Ed.). *Education policy in developing countries* (pp. 13–64). (edited by). Chicago, IL: University of Chicago Press.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies, 64*(4), 605–654.

Imbens, G. W. (2015). Matching methods in practice: three examples. *Journal of Human Resources, 50*(2), 373–419.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science, 340*(6130), 297–300.

Lockheed, M., Harris, A., & Jayasundera, T. (2010). School improvement plans and student learning in Jamaica. *International Journal of Educational Development, 30*, 54–66.

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

Min, Y., Yanqing, D., & Wenbin, H. (2012). *Southwest basic education project (SBEP): analysis of the impact of SBEP on student achievement.* Cambridge, United Kingdom: Cambridge Education.

Murnane, R. J., & Ganimian, A. J. (2014). *Improving educational outcomes in developing countries: lessons from rigorous evaluations.* NBER Working Paper Series No. 20284. Cambridge, MA.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics, 161*(2), 203–207.

Paqueo, V., & Lopez-Acevedo, G. (2003). *Supply-side school improvement and the learning achievement of the poorest children in indigenous and rural schools: the case of pare.* Washington, DC: The World Bank.

Pritchett, L. (2013). *The rebirth of education: schooling ain't learning.* Washington, DC: Center for Global Development.

Rodríguez, C., Sánchez, F., & Armenta, A. (2010). Do interventions at school level improve educational outcomes? Evidence from a rural program in Colombia. *World Development, 38*(3), 415–428.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., & Eyers, J.et al. (2015). Interventions for improving learning outcomes and access to education in low- and middle-income countries: A systematic review: International initiative for impact evaluation (3ie).

Tavares, P. A. (2015). The impact of school management practices on educational performance: Evidence from public schools in São Paulo. *Economics of Education Review, 48*, 1–15.