Great Things Come to Those Who Wait: Experimental Evidence on Performance-Management Tools and Training in Public Schools in Argentina[*]

Rafael de Hoyos[†]
World Bank

Alejandro J. Ganimian[‡]
New York University

Peter A. Holland[§]
World Bank

September 2, 2020

Abstract

In recent years, several studies have found that informing primary schools of their students' achievement leads to changes in school management, instruction, and learning. We conducted an experiment in Salta, Argentina to understand whether school systems should go one step further and support principals to act on the information they receive. We randomly assigned 100 public primary schools to: a diagnostic-feedback group, in which we administered math and reading tests and made results available to principals; or a performance-management group, in which we also provided principals with training and an online dashboard to develop, implement, and monitor school-improvement plans. The intervention had limited impact on students' performance in school during the study, but in the two years after it concluded, it reduced repetition rates in all target grades and it increased passing rates (reduced failure rates) for cohorts with two years of exposure. In fact, when we compare the schools in our study to other urban and semi-urban schools, they have lower dropout rates both during and after the study across all target grades. Our study suggests that school-management practices take a longer time to change than typically expected and highlight the importance of tracking post-intervention outcomes.

JEL codes: C93, I21, I22, I25.

Keywords: Diagnostic feedback, performance management, student assessments, Argentina.

[*] We gratefully acknowledge funding provided by the World Bank for this study. We thank Alfonsina Barraza, Analía Berruezo, Patricia Di Pasquale, Roberto Dib Ashur, Ximena Fernández, and Rosana Hernández at the Ministry of Education, Science, and Technology of Salta, and Eduardo Cascallar, Jorge Fasce, and Gustavo Iaies at the Center for Studies of Public Policies for making this study possible. We also thank Nicolás Buchbinder, María Cortelezzi, Alvina Erman, Andrés Felipe Pérez, María José Vargas, and especially Maya Escueta, who provided excellent research assistance. This study was registered with the AEA Trial Registry (RCT ID: AEARCTR-0002453). All views expressed are those of the authors and not of any of the institutions with which they are affiliated.

[†] Lead Economist, Education, World Bank. E-mail: `rdehoyos@worldbank.org`.

[‡] Assistant Professor of Applied Psychology and Economics, New York University Steinhardt School of Culture, Education, and Human Development. E-mail: `alejandro.ganimian@nyu.edu`.

[§] Lead Education Specialist, World Bank. E-mail: `pholland@worldbank.org`.

1 Introduction

There is growing evidence that the quality of school management matters for student learning. The bulk of research on this question has been conducted in the United States and Canada, where studies have found that school value-added estimates (i.e., year-on-year differences in student achievement, adjusted for school and student characteristics) vary across principals, across schools, and within schools over time (see, e.g., Branch, Hanushek and Rivkin 2012; Dhuey and Smith 2014; Grissom, Kalogrides and Loeb 2015), and where management metrics (e.g., operations, targets, and human-resource practices) are predictive of student achievement (Bloom et al. 2015). In recent years, these measures have been adapted for developing countries (see, e.g., Crawfurd 2017; Leaver, Lemos and Scur 2018; Lemos and Scur 2016).

Yet, in spite of the growing consensus on the importance of school management, there is very little evidence on effective interventions to improve it in low- and middle-income countries. Until recently, most randomized evaluations had focused on three main types of initiatives: (a) increasing parent and/or community pressure on schools to improve learning outcomes (see, e.g., Banerjee et al. 2010; Barr et al. 2012; Pandey, Goyal and Sundararaman 2009); (b) providing schools with grants (e.g., Beasley and Huillery 2016; Blimpo, Evans and Lahire 2011; Carneiro et al. 2015; Gertler, Patrinos and Rubio-Codina 2012; Mbiti et al. 2019; Pradhan et al. 2014); and (c) subsidizing the demand for or supply of privately schools (e.g., Alderman, Kim and Orazem 2003; Angrist, Bettinger and Kremer 2006; Angrist et al. 2002; Barrera-Osorio and Raju 2015; Barrera-Osorio et al. 2013; Bettinger, Kremer and Saavedra 2010; Muralidharan and Sundararaman 2015; Romero, Sandefur and Sandholtz 2017; Wong et al. 2013). Notably, however, none of these initiatives seeks to improve the *capacity* of public-school principals.[1]

One approach to improve school management that has had promising results in upper-middle income countries is providing principals with information on the achievement of their students. In the Province of La Rioja, Argentina, de Hoyos, Ganimian and Holland (2019) found that offering public primary schools reports that analyzed the performance of their students on standardized tests led principals to use achievement data to inform management decisions, teachers to assign more work during and after school, and students to improve their scores not only on the tests on which the reports were based but also on the national assessment. A related initiative in Mexico had similar results (de Hoyos, García-Moreno and Patrinos 2017).[2]

The encouraging findings from these studies raise the question of whether school systems should go one step further and attempt to help principals act on the feedback they receive.

[1] One notable exception is an ongoing evaluation of an intervention to increase principals' role in supporting and supervising differentiated instruction in Ghana (see Beg et al. 2019).

[2] In our study, we discuss at some length why we believe that this type of intervention is better suited for upper-middle income countries, where most teachers attend school, than for lower-middle income countries, where a considerable share of teachers is regularly absent (see Muralidharan and Sundararaman 2010).

School management capacity in developing countries is low (Bruns, Filmer and Patrinos 2011; Mbiti 2016; World Bank 2018), so principals may stand to benefit from additional assistance. Yet, if poor school management is symptomatic of broader problems in the public sector (Adelman et al. 2018; Andrews, Pritchett and Woolcock 2017; Finan, Olken and Pande 2017), governments may be ill-equipped to offer this support. Maybe they *should*, but they cannot.

This paper presents one of the first experimental evaluations of an initiative to support principals to make management decisions based on assessment results in a developing country. We randomly assigned 100 public primary schools in the Province of Salta, Argentina, to: a "diagnostic-feedback" (control) group, in which we administered math and reading tests to grades 3 to 5 and made results available to schools through user-friendly reports for two years (as in the intervention that we had evaluated in de Hoyos, Ganimian and Holland 2019); or a "performance-management" (treatment) group, in which we did the same and also provided principals with professional-development workshops and access to an online dashboard to design, implement, and monitor the implementation of their own school-improvement plans. We evaluate the effect of performance management, over and above that of diagnostic feedback.

We report three main sets of results based on this experiment. First, performance management had limited effect on students' performance in school during the two years of the intervention. When we estimate its impact across all grades targeted by the intervention, we find increases in the share of students who passed the grade of 1 percentage point (pp.) in 2015 and 2016 (over a control mean of 96%) and commensurate reductions in failure rates, but these effects are only statistically significant when we account for pre-intervention levels of those indicators. When we perform the estimation separately for each grade, we only find two isolated impacts: a 0.18 pp. reduction in dropout rates in grade 3 in 2015 and a 1.4 pp. reduction in failure rates in grade 4 in 2016. We do not find any statistically significant effects on student achievement. Consistent with these null effects, we also find no evidence of changes in teachers' attendance, punctuality, or time allocation, or in principals' involvement in instructional leadership.

Second, the effects of performance management emerged in the two years after its conclusion. When we estimate its impact across all target grades, we find increases in passing rates of between 2.2 and 2.3 pp. in 2017 and of 1.7 pp. in 2018, commensurate reductions in failure rates, and reductions in repetition rates of 2.9 to 3 pp. in 2017 and of 1.5 to 2.1 pp. in 2018. Reassuringly, when we estimate effects separately by grade, we see that they are driven by students and teachers who had more exposure to the performance-management intervention. Grade 3 students, whose teachers received two years of the intervention, but who were not exposed, saw decreases in repetition rates of 2.3 to 2.5 pp. in 2017 and 2.2 to 2.3 pp. in 2018. Grade 4 students, whose teachers received the intervention for a year, saw increases in passing rates of 1.6 pp. and decreases in repetition rates of 3.3 to 3.5 pp. if they also received it for a year (fourth-graders in 2017), and no effects if they had no exposure (fourth-graders in 2018).

Grade 5 students, whose teachers received the intervention for a year, saw increases in passing rates of 2.2 pp. and reductions in repetition rates of 2.7 to 3.3 pp. if they did not receive it (fifth-graders in 2017), and increases in passing rates of 2.9 pp. and reductions in repetition rates of 2.3 to 2.4 pp. if they also received the intervention for a year (fifth-graders in 2018). These effects represent sizable reductions in repetition rates with respect to the control group: in some grades and years, they amount to reducing repetition rates by 43% to 63%.

Lastly, the added value of performance management over diagnostic feedback is remarkable, given that the latter already constitutes an improvement upon business-as-usual conditions. When we compare treatment and control schools to all other urban and semi-urban schools, we observe that all schools in our study have lower dropout rates than out-of-sample schools. Diagnostic feedback reduced these rates by 0.49 and 0.52 pp. during the intervention period, in 2015 and 2016, and by 0.85 and 0.74 pp. in the two years that followed, 2017 and 2018. These effects are statistically significant when we estimate them pooling across all grades and when we estimate them for each target grade, both during and after the intervention period. They amount to reductions of 59% to 78% with respect to the dropouts in non-study schools. Performance management also has a statistically significant negative effect on dropouts after the intervention period, regardless of whether we estimate them across or within target grades, and we cannot rule out the possibility that both interventions had equal effects on dropouts. Yet, performance-management schools also had higher passing rates and lower repetition rates than out-of-sample schools, and for those outcomes we can rule out equal treatment effects. To our knowledge, these represent the largest effects on grade repetition and school dropouts of any education intervention previously evaluated in Argentina, the context of our study.

Our study makes four contributions to research on school management in developing nations. First, they add to mounting experimental evidence indicating that informing schools of their students' achievement and helping them act on that information improves their performance (see de Hoyos, Ganimian and Holland 2019; de Hoyos, García-Moreno and Patrinos 2017). Second, they demonstrate that school-leadership interventions take time to change practices, both in management and instruction, and ultimately students' daily experiences at school. This insight might help explain the prevalence of null results in this literature (e.g., Ganimian and Freel 2020; Muralidharan and Singh 2020; Muralidharan and Sundararaman 2010). Third, our study illustrates the benefits of combining primary and administrative data whenever possible to both track effectiveness beyond the intervention period and avoid social desirability. Finally, it contributes to evidence of the disconnect between school performance and learning in the developing world (e.g., Ganimian et al. 2020; Muralidharan, Singh and Ganimian 2019).

The rest of the paper is structured as follows. Section 2 describes the context, interventions, sampling, and randomization. Section 3 presents the data. Section 4 discusses the empirical strategy. Section 5 reports the results. Section 6 discusses implications for policy and research.

2 Experiment

2.1 Context

Schooling in Argentina is compulsory and free from age 4 until the end of secondary school. In 12 of the 24 provinces including Salta, primary education runs from grades 1 to 7 and secondary education from grades 8 to 12 (DIEE 2020).[3] The Argentine school system serves 11.5 million students: 1.85 million in pre-primary education, 4.83 million in primary education, 3.87 million in secondary education, and over 980,000 in tertiary education (DIEE 2020).[4] The school year runs from February to December, but the start and end dates vary across provinces.

According to the National Education Law of 2006, each of the 24 sub-national (province) governments in Argentina is responsible for providing pre-primary, primary, and secondary education to its inhabitants, and the national government is responsible for higher education as well as technical and financial assistance to the provinces (National Education Law 2006). Since 1993, the Ministry of Education at the national level has been in charge of administering the national assessment of student achievement (formerly known as the *Operativo Nacional de Evaluación* and currently as *Aprender*) in coordination with its province-level counterparts.

Most primary-school aged children in Argentina are enrolled in school. According to the latest internationally comparable data available, the country's net primary enrollment rate is 99%, and nearly all students who complete this level go on to secondary school (UNESCO 2020). Yet, many primary-school graduates still struggle to reach minimum levels of academic skills: in the 2018 *Aprender*, 25% of sixth-graders performed in the lowest two of the four proficiency levels in reading ("basic" and "below basic") and 43% did so in math (SEE-MEDN 2019*b*). Multiple changes in the design and administration of national and regional assessments have rendered comparisons of the performance of primary-school students over time challenging.[5]

Argentina is an interesting setting to study the effects of performance management training and tools for public schools. From 2000 to 2015, the federal government took multiple steps to limit the generation, dissemination, and use of student achievement data, including: reducing the frequency of its national assessment (first from every year to every two years, and then to every three years), suspending the publication of results at the province level (only making

[3] In the other 12 provinces, primary runs from grades 1 to 6 and secondary from grades 7 to 12.

[4] These figures only refer to common education and exclude special and adult education.

[5] If we compare the results of the 2018 installment of *Aprender* to those of 2016 (the first year in which it was administered), sixth-graders improved in reading but did not make progress in math (SEE-MEDN 2017). Comparisons to earlier installments are problematic due to changes in the test's content and methodology. Similarly, if we compare the 2013 assessment of Latin American and Caribbean school systems to that of 2006, Argentine third- and sixth-graders have improved in math, but not in reading (UNESCO-LLECE 2014). Yet, comparisons to its first installment in 1997 are not possible due to changes in several aspects of the test (for a more detailed discussion of these issues, see Ganimian 2009, 2014, 2015*b*).

results available by geographic region), and prohibiting the public disclosure of results at the school, teacher, and student level in the 2006 National Education Law (see Ganimian 2015*a*).[6] Some of these steps were temporarily reversed from 2017 to 2019, when a new administration began notifying each school of its students' performance on the national assessment, but it was voted out of office in 2020, and since then the continuity of this strategy has been uncertain. Therefore, in spite of having a long-standing assessment, Argentina has traditionally provided principals with little to no information on the academic skills of the students at their schools. This is why any efforts to provide such data to schools are likely to be more impactful in this setting than they would be in similar countries with a steadier information flow to schools.

We conducted our study in Salta for two main reasons. First, it is one of the lowest performing provinces in the national assessment, and thus stands to benefit from interventions that seek to improve learning outcomes: in the 2018 *Aprender*, 25% of sixth-graders scored in the lowest two of the four proficiency levels in reading and 40% did so in math (SEE-MEDN 2019*a*).[7] Second, it was one of the few provinces with the political will to experiment with a sub-national assessment. At the time, the test was endorsed by the governor and the education minister.

2.2 Sample

The sampling frame for the study included all 397 public primary schools located in urban and semi-urban areas of Salta.[8] We arrived at this frame as follows. First, of the 838 primary schools in the province, we excluded all 85 private schools because we were interested in the effect of the intervention on public schools. Then, we dropped all 481 public schools in rural areas because they are spread across the province, which would have limited our capacity to implement the intervention.[9] Next, we dropped 16 schools that did not have the data we needed to select our sample. We drew a random sample of 100 urban and semi-urban public primary schools from this frame, stratified by geographic area (i.e., urban and semi-urban).

The schools in our sample are comparable to all public primary schools in the province, and even more so to other urban and semi-urban public primary schools (Table A.1, Appendix A). The average school in our sample enrolls more students, but this is mostly because we excluded

[6] Notably, these policies stood in stark contest with those of other middle- or upper-middle income countries in South America (e.g., Brazil, Chile, Colombia, and Peru), which have technically robust and long-standing assessments and use them for multiple purposes (Ferrer 2006; Ferrer and Fiszbein 2015).

[7] These figures resemble the national averages (reported earlier in this section), but those averages are driven by the Province of Buenos Aires, which serves about a third of the country's students.

[8] Throughout this paper, we use the terms "semi-urban" to refer to areas locally known as *rurales aglomeradas* and "rural" for areas known as *rurales dispersas*.

[9] Note, however, that while rural schools account for a large share of the total number of public schools in the province (64% of the total), they serve a small share of students (about 18% of the total).

rural schools, which are typically smaller. We we only find a few isolated impacts (mostly in grade 5), which are likely to have emerged by chance due to multiple-hypothesis testing.

We sampled students and teachers to obtain *cross-sectional* information in grade 3 every year, as well as *longitudinal* information on the students who started grade 3 in 2014. Thus, in 2014, all students and teachers from grade 3 participated; in 2015, all students and teachers from grades 3 and 4 participated; and in 2016, all students and teachers from grades 3 and 5 participated. All principals in selected schools participated in the study.

2.3 Randomization

We randomly assigned the 100 schools in our sample to: a "diagnostic feedback" (control) group, in which we administered standardized tests of math and reading and made the results available to schools through user-friendly reports for two years (2015 and 2016); or a "performance management" (treatment) group, in which we also provided principals with professional-development workshops and access to an online dashboard to design, implement, and monitor school-improvement plans. We stratified our randomization by geographic area (i.e., urban and semi-urban) to increase statistical power. This setup allows us to evaluate the effect of performance management, over and above that of diagnostic feedback.

2.4 Intervention

Table 1 shows the timeline for the interventions and rounds of data collection for the study. The school year in Argentina starts in February and ends in December. As the table shows, we administered the student assessments at the end of each year and we delivered school reports based on the prior-year assessments at the start of the following year. We granted access to dashboards and conducted workshops for school supervisors and principals during each year.

2.4.1 Diagnostic-feedback (control) group

The diagnostic feedback intervention provided schools with reliable, timely, and actionable data on student learning outcomes to inform school management and classroom instruction. At the beginning of each year, schools randomly assigned to the control group received reports that summarized the results of student assessments administered at the end of the prior year.[10]

The reports were brief (10 pages) and had four sections: (a) an introduction, which described the assessments and reported the percentage of students at the school who completed them;

[10] We specify the grades assessed on each year in Table 1.

(b) an overview of the school's average performance, which included the school's average score in each grade and subject, the change in each score from the previous year, and comparisons between the school's scores and those of the average school in the area and in the province;[11] (c) an analysis of the distribution of the school's performance, which included a box-and-whiskers plot for the school and the province for each grade and subject; and (d) a "traffic light" display of the school's performance on each item of the assessments for each grade and subject.[12]

2.4.2 Performance management (treatment) group

The performance management intervention also provided school supervisors and principals with support to use student achievement data to develop a school-improvement plan and monitor its implementation. At the beginning of each year, schools randomly assigned to the treatment group received the same reports as the control group. During the school year, and for two consecutive years, these schools also received 11 workshops (six in 2015 and five in 2016), and access to dashboards to track progress on their school-improvement plan.

The workshops in both years asked principals to develop a school-improvement plan (i.e., a plan to improve one or more aspects of their school), upload information on that plan to an online dashboard, and use the dashboard to monitor the implementation of the plan. In 2015, the workshops also covered how to conduct classroom observations and give teachers feedback. In 2016, the workshops also focused on effective teaching practices in math and language.[13]

The dashboards had three main sections: (a) an overview that presented each school's average performance in the assessments, passing rates in both subjects, and student absenteeism rates; (b) a section that described the school's progress towards the goals in its school-improvement plan (e.g., targets for classroom observations, parent-teacher meetings, supervisor-principal meetings, or principal-teacher meetings); and (c) a section that reported on the school's internal efficiency (e.g., enrollment, passing, repetition, and overage rates).[14]

As Table 1 shows, workshop participation and dashboard use in 2016 were lower than in 2015. In section 5.2, we use the endline data to discuss variation in dosage among treatment schools.

3 Data

As Table 1 indicates, we administered student assessments of math and reading and surveys of students, teachers, and principals in control and treatment schools on each year of the study

[11] All scores were scaled and linked using a two-parameter Item Response Theory model.

[12] A template of the report can be accessed at: `https://bit.ly/2xCPKbq`.

[13] The workshops were based on the recommendations in Boudett, City and Murnane (2005).

[14] A template of the dashboard in English can be found at: `http://bit.ly/2cYBXOR`.

(from 2014 to 2016). We also obtained access to data on students' performance in school from the annual census of schools (for 2014 to 2018) and on students' achievement from the national assessment (for 2016) for all schools in the province.

3.1 Independent assessments

We administered student assessments of math and reading before the intervention (in 2014) and after one and two years (in 2015 and 2016) in both control and treatment schools.[15] The assessments evaluated what students ought to know and be able to do according to: (a) the national curriculum (*Contenidos Básicos Comunes*); (b) the topics of the curriculum that the national government has identified as priorities (*Núcleos de Aprendizaje Prioritario*); and (c) the curriculum of the province (*Diseño Curricular de Salta*). Specifically, the math assessment covered four content domains (numbers, geometry, measurement, and probability and statistics) and four cognitive domains (identifying mathematical concepts, understanding and using symbolic math, performing calculations, and solving abstract and applied problems). The reading assessment covered three content domains (narrative, informative, and short texts) and four cognitive domains (locating information, understanding relationships between parts of texts, identifying the main idea of texts, and interpreting the meaning of words). Each assessment included 30 to 35 multiple-choice items.

We used a two-parameter Item Response Theory (IRT) model to scale the assessment results in a way that accounts for differences between items (their difficulty and capacity to distinguish between students of similar ability) and leverages common items to link results over time.[16] Appendix B provides further details on the design, scaling, and linking of the assessments, as well as the distribution of scores for all subjects, grades, and years of the study.

3.2 Student surveys

We also administered surveys of students. In 2014 (i.e., the year before the intervention), the surveys asked about students' demographic characteristics, home assets, schooling trajectory, and study supports to allow us to describe our study sample. In 2015 and 2016 (i.e., the years of the intervention), they asked students about their teachers' effort, as measured by the frequency of attendance, punctuality, and a set of classroom activities, and about their

[15] Some students may be missing test scores because they were absent on the day of the assessments, they were present but excused from the tests, they dropped out of school, or they transferred to another school. However, we find no evidence of treatment schools having more students than control schools on such days. We do not have school-level data on transfers or students who were present and excluded on test day, but we have no reason to believe that either occurred frequently or differentially across groups.

[16] The assessments are available at: `https://bit.ly/2vp1AoQ` (2014), `https://bit.ly/2TSEMqU` (2015), and `https://bit.ly/3d3gmTh` (2016).

teachers' effectiveness, as measured by an abridged version of the Tripod survey developed by Ron Ferguson at Harvard (see, e.g., Ferguson 2010, 2012; Ferguson and Danielson 2014).[17]

3.3 Teacher surveys

We also conducted surveys of teachers. In 2014, we asked about demographic characteristics, education and experience, professional development, and teaching practices to describe our sample. In 2015 and 2016, we asked about aspects that could plausibly be influenced by the intervention (e.g., monitoring and evaluation practices at their schools and job satisfaction).[18]

3.4 Principal surveys

We administered surveys of principals. In 2014, we asked about demographic characteristics, education and experience, professional development, and school infrastructure to describe our sample. In 2015 and 2016, we asked about aspects that could be affected by the intervention (e.g., management practices and resources and materials).[19]

3.5 National assessment

We also gained access to the results of the 2016 national assessment of sixth-graders, which was administered a year after the end of the intervention.[20] This assessment evaluated the math and reading skills of students who were in grade 4 in 2014 and who were exposed to the intervention for one year. We use these data to estimate the effect of the intervention.

3.6 Census of schools

Finally, we also obtained access to the data on students' performance (e.g., passing, failure, repetition, and dropout rates) collected annually through the census of schools. We use the 2014 data to compare in- and out-of-sample schools and to check balance between experimental groups and we use the data for 2015 to 2018 to estimate the effect of the intervention. Importantly, these data are reported at the grade and school levels, so all impact estimations on these indicators are conducted at those levels (rather than at the student level).

[17] The surveys are at: `https://bit.ly/3b2oqSe` (2014) and `https://bit.ly/38WrQ7E` (2015-2016).
[18] The surveys are at: `https://bit.ly/39VJN7S` (2014) and `https://bit.ly/2WithdW` (2015-2016).
[19] The surveys are at: `https://bit.ly/2Wm48iD` (2014) and `https://bit.ly/2IQvWnp` (2015-2016).
[20] There is an assessment for third-graders, but it is only administered to a sample of schools.

4 Empirical strategy

4.1 Added value of performance management

We estimate the intent-to-treat or ITT effect of performance management training and tools, over and above that of diagnostic feedback, after one and two years, by fitting the following model for the 100 public primary schools in our study:

$$Y_{igs}^{t} = \alpha_{r(s)} + \bar{Y}_{gs}^{t=0}\psi + T_s\beta + \epsilon_{igs}^{t} \tag{1}$$

where Y_{igs}^{t} is the outcome for student i in grade g and school s for year t, $r(s)$ is the randomization stratum of school s and $\alpha_{r(s)}$ its fixed effect, $\bar{Y}_{gs}^{t=0}$ is the school- or grade-level mean of the outcome at baseline if available,[21] and T_s is an indicator variable that equals 0 for control (diagnostic feedback) schools and 1 for treatment (performance management) schools. The parameter of interest is β, which captures the added value of performance management. We use cluster-robust standard errors to account for within-school correlations across students in outcomes. We also test the sensitivity of our estimates to the inclusion of $\bar{Y}_{gs}^{t=0}$.

We also fit several variations of this model, including: (a) one in which outcomes are measured at the teacher or principal level (to estimate the impact of performance management on classroom instruction and school management); (b) one in which outcomes are measured one and two years after the end of the study (to estimate fadeout and/or dormant effects);[22] and (c) one in which we interact the treatment dummy with student-level covariates (to estimate the differential effect of the intervention on sub-groups of students).

4.2 Effect of diagnostic feedback and performance management

We leverage the data we have for some outcomes for *all* schools in Salta to estimate the ITT effect of diagnostic feedback and performance management over business-as-usual practices by fitting the following model for public primary schools in urban and semi-urban areas:[23]

$$Y_{igs}^{t} = \phi_{a(s)} + \bar{Y}_{gs}^{t=0}\delta + I_s'\gamma + \epsilon_{igs}^{t} \tag{2}$$

[21] We have baseline measures of students' performance in school and students' achievement on independently administered assessments (see section 3). Yet, we cannot account for these baseline measures at the student level because students were not assigned unique identifiers to allow us to track them over time.

[22] We observe students' performance in school for two years after the interventions ended (see section 3).

[23] We observe students' performance in school and achievement in the national assessment (see section 3).

where $a(s)$ is the geographic area of school s and $\phi_{a(s)}$ its corresponding fixed effect,[24] I'_s is a vector of intervention indicators, and everything else is defined as above. The parameters of interest are the elements of γ, which capture the effect of each intervention. Again, we use cluster-robust standard errors and test the sensitivity of our estimates to baseline adjustments.

5 Results

5.1 Balancing checks

Control and treatment schools were similar at baseline. This holds true regardless of whether we compare them on administrative data on students' performance in school (Table 2), the independent assessments we administered at the start of the study (Table A.2), or background variables from the surveys of students, teachers, and principals at that time (Tables A.3-A.5). We only find a few small and isolated statistically significant differences, which we believe are likely to have emerged due to chance in light of the multiple hypotheses that we are testing. All other differences are consistently and precisely estimated around zero.

5.2 Implementation fidelity

The interventions were implemented mostly as intended. First, as Table 1 indicates, nearly all schools participated in the student assessments that we used to prepare the school reports. Second, as the table also shows, the vast majority of schools had at least one representative (a school supervisor or principal) attend the workshops to develop, monitor, and implement school-improvement plans (except for workshop 8, for which only 18% of principals attended). Yet, implementation differed across years. Attendance to workshops and the use of dashboards were higher in 2015, and some workshops in 2016 were attended by small shares of principals.[25]

5.3 ITT effects during the intervention period

Performance management training and tools added little value to diagnostic feedback during the two years in which the intervention was implemented in schools (i.e., from 2015 to 2016).

[24] As we discuss in section 2.3, we stratified random assignment of schools to groups by geographic area, so this is equivalent to including randomization fixed effects for all schools in this estimation.

[25] We also asked principals about implementation fidelity, but their responses were problematic (Table A.6). Specifically, many principals in control schools reported engaging in activities reserved for the treatment group (e.g., attending workshops, accessing the online portal), even if implementation data (e.g., attendance records to workshops or logins to the portal) indicate that this was not the case. We suspect that this confusion arose because, during our study, many schools were participating in other school-management programs (Table A.7). For these reasons, we rely on the administrative data on implementation reported above.

When we estimate its impact on students' school performance across target grades (3 to 5), we find effects on passing and failure rates, but they are only statistically significant when we account for pre-intervention school averages of those indicators (Table 3, panels A and B).[26] When we estimate effects for each grade targeted by the intervention separately, we only find two statistically significant impacts: a 0.2 pp. reduction in dropout rates in grade 3 in 2015 and a 1.4 pp. reduction in failure rates in grade 4 in 2016 (Tables A.8-A.10, panels A and B).

We do not find any statistically significant effects on any other outcomes during this period. Students in treatment schools performed on par with their control peers on the independent assessments of math and reading that we administered in both 2015 and 2016 (Table 4). This is also true when we estimate treatment effects separately by grade (Tables A.11-A.13).

Consistent with these results, we find little evidence of impact on two potential mechanisms: classroom instruction and school management practices. In the two years of the intervention, teachers in treatment schools were no more likely to attend school, arrive on time, or allocate lesson and/or school time differently than their counterparts in control schools (Table A.14). They were no more likely to exert more effort inside the classroom either (Tables A.15-A.16). They were not rated more favorably by their students on a frequently-used scale (Table A.17). Principals in treatment schools do report being more likely to share the results of student assessments with parents and the school community than those in control schools (Table A.18), but this does not seem to increase their involvement in instructional leadership (e.g., observing lessons, giving teachers feedback, or leading demonstration lessons or trainings, Table A.19). Overall, all results during the study period suggest that the intervention had a limited effect.

5.4 ITT effects after the intervention period

The effect of performance management begin to emerge in the two years after the intervention. When we estimate its added value over diagnostic feedback across all target grades (3 to 5), we find statistically significant increases in passing rates of between 2.2 and 2.3 pp. in 2017 and of 1.7 pp. in 2018, commensurate reductions in failure rates, and reductions in repetition rates of 2.9 to 3 pp. in 2017 and of 1.5 to 2.1 pp. in 2018 (see Table 3, panels C and D).

Reassuringly, when we estimate effects separately by grade, we see that they are driven by students and teachers who had more exposure to the performance-management intervention. The cohorts of students who were in third grade in 2017 and 2018, whose teachers received the intervention for two years, but who were not exposed to the interventions themselves, saw decreases in repetition rates of 2.3 to 2.5 pp. in 2017 and of 2.2 to 2.3 pp. in 2018 (Table A.8). The students who were in grade 4 in 2017, who were exposed to the intervention for one year

[26] As we mention in section 4, this information is reported at the grade and school levels, so we estimate effects at those levels, with or without accounting for the corresponding pre-intervention averages.

(the same time as their teachers), saw increases in passing rates of 1.6 pp., decreases in failure rates of the same magnitude, and reductions in repetition rates of 3.3 to 3.5 pp. (Table A.9). Those who were in grade 4 in 2018, who had teachers with a year of exposure but who did not receive it themselves, only saw a marginally significant reduction in failure rates in 2018. The students who were in grade 5 in 2017 and 2018, who had zero and one years of exposure, respectively, and whose teachers were exposed to the intervention for a year, saw increases in passing rates of 2.2 pp. in 2017 and 2.9 pp. in 2018, similar decreases in failure rates, and reductions in repetition rates of 2.7 to 3 pp. in 2017 and of 2.3 to 2.4 pp. in 2018 (Table A.10). These effects represent sizable reductions in repetition rates with respect to the control group: in some grades and years, they amount to reducing repetition rates by 43% to 63%.

We do not find evidence that these improvements in school performance led to more learning. When we estimate the effect of the intervention on students who were in sixth grade in 2016, we do not observe any statistically significant effects in the national assessment (Table 5).[27] The difference between control and treatment schools is about 0.01 standard deviations (SDs) in both math and reading, and it is not statistically significant in either of these subjects. Yet, based on the 95% confidence intervals, we cannot rule out effects of up to 0.22 SDs, so the intervention may have had effects on learning and we are under-powered to detect them.

5.5 Comparison of in- and out-of-sample schools

As we mention in sections 3.5 and 3.6, we have data from the annual school census and national assessment for all public primary schools in Salta, not just for the 100 schools in our sample. As we discuss in section 4.2, this allows us to estimate the effect of diagnostic feedback and performance management with respect to all out-of-sample urban and semi-urban schools.[28]

When we do so, we find that diagnostic feedback reduced dropout rates during and after the intervention period, making the effects of performance management all the more remarkable. When we estimate effects across all target grades, we find diagnostic feedback reduced dropout rates by 0.49 pp. in 2015, 0.52 pp. in 2016, 0.85 pp. in 2017, and 0.74 pp. in 2018 (Table 6). Given that the out-of-sample mean dropout rates were 0.8% in 2015, 0.7% in 2016, 1.1% in 2017, and 1.2%, these effects amount to reductions in dropout rates of between 59% and 78%. Performance management also led to statistically significant decreases in dropout rates of 0.59 pp. (71%) in 2015, 0.37 pp. (50%) in 2016, 0.86 pp. (79%) in 2017, and 1 pp. (81%) in 2018.[29] These results indicate that, while performance management may not have reduced dropouts

[27] As we explain in section 3, this is the only grade for which the national assessment was census-based (i.e., it is supposed to cover all students in a grade, rather than a sample of them).

[28] As we explain in section 2.2, our sample is composed entirely of urban and semi-urban schools.

[29] Unfortunately, however, we do not have sufficient statistical power to detect differences in treatment effects on dropout rates between experimental groups.

with respect to diagnostic feedback (see section 5.4), it did so with respect to business-as-usual conditions in out-of-sample public primary schools in urban and semi-urban areas in Salta. This pattern of results makes sense, given that diagnostic feedback is part of the performance management intervention (see section 2.4) and dropout rates in primary schools are very low.

Performance management also had positive impacts on other school-performance indicators. It increased passing rates on every year during and after the intervention period: by 1.3 pp. in 2015, 1.5 pp. in 2016, 2.8 pp. in 2017, and 1.7 pp. in 2018 (between 1% and 2%; Table 6).[30] Consistent with our experimental results (section 5.4), it also reduced repetition rates in both years of the post-intervention period: by 1.2 pp. (30%) in 2017 and 1.4 pp. (40%) in 2018. These results confirm performance management improved upon business-as-usual conditions, and they are aligned with those from our estimations by target grade (Tables A.20- A.22).[31]

Neither diagnostic-feedback, nor performance-management schools had better results on the national assessment than out-of-sample public primary schools in urban and semi-urban areas (Table 7). Yet, this may be potentially explained in part by their effects on dropout rates. If these interventions prevented low-performing students from dropping out, it is possible that they improved the learning outcomes of their peers, but that their gains are not reflected in comparisons of mean scores between in- and out-of-sample schools (because the scores of the retained low performers bring the average score of in-sample schools down, while the average score of out-of-sample schools appear to improve because their low performers dropped out).

6 Conclusion

We present experimental evidence on the impact of performance-management for public primary schools in Salta, Argentina and find that combining low-stakes information on student achievement with support to act on it radically improves students' performance in school, though these results take longer to emerge than the typical randomized evaluation may assume. In spite of having a comprehensive measurement strategy that tracked student learning as well as potential mechanisms (e.g., classroom instruction and school management) for two years, we did not start seeing the effects of the intervention until after the intervention had concluded. Once they emerged, however, treatment effects on school performance were transformative:

[30] These increases in passing rates are not always accompanied by statistically significant decreases in failure rates, but that is partly because average failure rates in out-of-sample schools are smaller in magnitude, and thus effects on them need to be estimated with greater precision to reach statistical significance. Reassuringly, however, differences in sign and magnitude are consistent with the treatment effects on passing rates.

[31] We do not observe these impacts for diagnostic-feedback schools, even if the performance-management intervention incorporates the diagnostic-feedback component (see section 2.4). We do not know why this is the case, but we conjecture that it could be because some schools need the additional support provided by the performance-management training and tools to make sense of the feedback they receive.

diagnostic feedback, by itself and when combined with performance management, halved dropout rates; and the combination also increased passing rates and reduced repetition rates. To our knowledge, these are the largest effects on students' performance in school to be found by any randomized evaluation in education in Argentina (see Ganimian and Murnane 2016).

Our study makes four contributions to research on school management in developing nations. First, they add to the growing body of experimental evidence suggesting that informing schools of their performance and helping them act on that information leads to improved outcomes—at least in low-information settings like Argentina, where such data are not already disclosed (see de Hoyos, Ganimian and Holland 2019; de Hoyos, García-Moreno and Patrinos 2017). This is a crucial implication for policy, given that such use of information is currently banned precisely in the same context where it has repeatedly been proven to work (see section 2.1). Yet, it is also important for research, given that prior failed attempts at combining feedback with capacity building may have led some to conclude that such efforts are not worth pursuing. Our study illustrates that the details of such initiatives are key to their effectiveness.

Second, perhaps the most important contribution of this study is to demonstrate that school leadership interventions take time to change management practices, instructional strategies, and ultimately students' performance in school—in fact, more time than we typically expect. This insight might help explain why so many such initiatives evaluated in recent years have yielded null results (see, e.g., Ganimian and Freel 2020; Muralidharan and Singh 2018; Muralidharan and Sundararaman 2010). Seeking to improve student outcomes by intervening in school management is a complex process that requires principals to improve their practices, teachers to make meaningful changes in their instruction, and students to respond positively. It seems reasonable that such processes demand some trial and error to produce results.

Third, and relatedly, our study demonstrates the benefits of complementing primary data (which is ubiquitous in experiments) with administrative data (which is far less common). If we had not done that, we would have incorrectly concluded the intervention was ineffective. Administrative data has multiple benefits over measures designed and collected by researchers. They minimize the risk of differential attrition across experimental groups, reduces the risk of social-desirability bias that emerges when implementers enquire beneficiaries on the same outcomes they seek to influence, and potentially even enhance the external validity of a study by allowing researchers to incorporate out-of-sample units into their analyses (provided that the government agencies that own these data authorize such uses; see, e.g., Ganimian 2020).

Finally, and more broadly, our study also contributes to growing evidence of the disconnect between students' performance in school and learning outcomes across the developing world. It resembles several other studies published in recent years that have found effects on student achievement, but not on school performance or vice versa (see, e.g., Ganimian et al. 2020;

Muralidharan, Singh and Ganimian 2019). The underlying causes of this trend are unclear: it may be partly that school grades capture not only academic, but also socio-emotional skills, as U.S.-based research suggests (see, e.g., Farrington et al. 2012; Jackson 2020; Kraft 2019), but it may also partly be that the grade-based expectations of school curricula characteristic of school exams may be inappropriate in contexts of wide heterogeneity in student preparation (see, e.g., Muralidharan and Zieleniak 2014; Pritchett and Beatty 2015). We would welcome further research into this important question in developing-country settings.

References

Adelman, M., R. Lemos, M. J. Vargas, and R. Nayar. 2018. "Managing for learning: (In)coherence in education systems in Latin America." *Unpublished manuscript.* Washington, DC: The World Bank.

Alderman, H., J. Kim, and P. F. Orazem. 2003. "Design, evaluation, and sustainability of private schools for the poor: The Pakistan urban and rural fellowship school experiments." *Economics of Education Review*, 22(3): 265–274.

Andrews, M., L. Pritchett, and M. Woolcock. 2017. *Building state capability: Evidence, analysis, action.* Oxford University Press.

Angrist, J., E. Bettinger, and M. Kremer. 2006. "Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia." *American Economic Review*, 96(3): 847–862.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer. 2002. "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment." *American Economic Review*, 92(5): 1535–1558.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani. 2010. "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India." *American Economic Journal: Economic Policy*, 2(1-30).

Barr, A., F. Mugisha, P. Serneels, and A. Zeitlin. 2012. "Information and collective action in community-based monitoring of schools: Field and lab experimental evidence from Uganda." *Unpublished manuscript.* Washington, DC: Georgetown University.

Barrera-Osorio, F., and D. Raju. 2015. "Evaluating the impact of public student subsidies on low-cost private schools in Pakistan." *The Journal of Development Studies*, 51(7): 808–825.

Barrera-Osorio, F., D. S. Blakeslee, M. Hoover, L. Linden, D. Raju, and S. Ryan. 2013. "Leveraging the private sector to improve primary school enrolment: Evidence from a randomized controlled trial in Pakistan." *Unpublished manuscript.* Cambridge, MA: Harvard Graduate School of Education (HGSE).

Beasley, E., and E. Huillery. 2016. "Willing but unable? Short-term experimental evidence on parent empowerment and school quality." *The World Bank Economic Review*, 31(2): 531–552.

Beg, S., A. Fitzpatrick, A. M. Lucas, E. Tsinigo, and H. Atimone. 2019. "Strengthening teacher accountability to reach all students (STARS)." (World Bank SIEF). Milestone 3: Observational survey field report.

Bettinger, E., M. Kremer, and J. E. Saavedra. 2010. "Are educational vouchers only redistributive?" *The Economic Journal*, 120(546).

Blimpo, M. P., D. K. Evans, and N. Lahire. 2011. "School-based management and educational outcomes: Lessons from a randomized field experiment." *Unpublished manuscript.*

Bloom, N., R. Lemos, R. Sadun, and J. Van Reenen. 2015. "Does management matter in schools?" *The Economic Journal*, 125(584): 647–674.

Boudett, K. P., E. A. City, and R. J. Murnane. 2005. *Data wise: A step-by-step guide to using assessment results to improve teaching and learning.* Cambridge, MA: Harvard Education Press.

Branch, G. F., E. A. Hanushek, and S. G. Rivkin. 2012. "Estimating the effect of leaders on public sector productivity: The case of school principals." (NBER Working Paper No. 17803). Cambridge, MA: National Bureau of Economic Research (NBER).

Bruns, B., D. Filmer, and H. A. Patrinos. 2011. *Making schools work: New evidence on accountability reforms.* Washington, DC: The World Bank.

Carneiro, P., O. Koussihouèdé, N. Lahire, C. Meghir, and C. Mommaerts. 2015. "Decentralizing education resources: School grants in Senegal." (NBER Working Paper No. 21063). Cambridge, MA: National Bureau of Economic Research (NBER).

Crawfurd, Lee. 2017. "School Management and Public–Private Partnerships in Uganda." *Journal of African Economies*, 26(5): 539–560.

de Hoyos, R., A. J. Ganimian, and P. A. Holland. 2019. "Teaching *with* the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina." *World Bank Economic Research.*

de Hoyos, R., V. A. García-Moreno, and H. A. Patrinos. 2017. "The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico." *Economics of Education Review*, 58: 123–140.

Dhuey, E., and J. Smith. 2014. "How important are school principals in the production of student achievement?" *Canadian Journal of Economics/Revue canadienne d'économique*, 47(2): 634–663.

DIEE. 2020. “Anuario estadístico 2019.” Buenos Aires, Argentina: Dirección de Investigación y Estadística Educativa (DIEE).

DiNIECE. 2009. “Estudio nacional de evaluación y consideraciones conceptuales: Educación primaria. Educación secundaria.” Ciudad Autónoma de Buenos Aires, Argentina: Subsecretaría de Planeamiento Educativo, Secretaría de Educación, Ministerio de Educación.

DiNIECE. 2012. “Operativo Nacional de Evaluación 2010: 3er y 6to año de la educación primaria. Informe de resultados.” Ciudad Autónoma de Buenos Aires, Argentina: Subsecretaría de Planeamiento Educativo, Secretaría de Educación, Ministerio de Educación.

Farrington, Camille A, Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W Johnson, and Nicole O Beechum. 2012. “Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance–A critical literature review.” Unpublished manuscript. The Unviersity of Chicago Consortium on Chicago School Research (CCSR). Chicago, IL.

Ferguson, R. F. 2010. “Student perceptions of teaching effectiveness.” Boston, MA: The National Center for Teacher Effectiveness and the Achievement Gap Initiative.

Ferguson, R. F. 2012. “Can student surveys measure teaching quality?” *Phi Delta Kappan*, 94(3): 24–28.

Ferguson, R. F., and C. Danielson. 2014. “How Framework for Teaching and Tripod 7Cs evidence distinguish key components of effective teaching.” T. J. Kane, K. A. Kerr & R. C. Pianta, *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: Jossey-Bass.

Ferrer, G. 2006. *Educational assessment systems in Latin America: Current practice and future challenges.* Partnership for Educational Revitalization in the Americas (PREAL).

Ferrer, G., and A. Fiszbein. 2015. “What has happened with learning assessment systems in Latin America? Lessons from the last decade of experience.” Washington, DC: The World Bank.

Finan, F., B. A. Olken, and R. Pande. 2017. “Handbook of field experiments.” , ed. A. V. Banerjee and E. Duflo Vol. II, Chapter 6: The personnel economics of the developing state. Oxford, UK: North Holland.

Ganimian, A. J. 2009. “How much are Latin American children learning? Highlights from the second regional student achievement test (SERCE).” Washington, DC: Partnership for Educational Revitalization in the Americas (PREAL).

Ganimian, A. J. 2014. “Avances y desafíos pendientes: Informe sobre el desempeño de Argentina en el Tercer Estudio Regional Comparativo y Explicativo (TERCE) del 2013.” Buenos Aires, Argentina: Proyecto Educar 2050.

Ganimian, A. J. 2015*a*. “El termómetro educativo: Informe sobre el desempeño de Argentina en los Operativos Nacionales de Evaluación (ONE) 2005-2013.” Buenos Aires, Argentina: Proyecto Educar 2050.

Ganimian, A. J. 2015*b*. “Pistas hechas en Latinoamérica: ¿Qué hicieron los países, escuelas y estudiantes con mejor desempeño en el Tercer Estudio Regional Comparativo y Explicativo (TERCE)?” Buenos Aires, Argentina: Red Latinoamericana por la Educación (Reduca) & Proyecto Educar 2050.

Ganimian, A. J. 2020. “Growth mindset interventions at scale: Experimental evidence from Argentina.” *Educational Evaluation and Policy Analysis*, 42(3): 417–438.

Ganimian, A. J., and R. J. Murnane. 2016. “Improving education in developing countries: Lessons from rigorous impact evaluations.” *Review of Educational Research*, XX(X): 1–37.

Ganimian, A. J., and S. H. Freel. 2020. “Can principal training improve school management? Short-term experimental evidence from Argentina.” *Unpublished manuscript.* New York, NY: New York University (NYU).

Ganimian, A. J., F. Barrera-Osorio, M. L. Biehl, M. Cortelezzi, and D. Valencia. 2020. “Hard cash and soft skills: Experimental evidence on combining scholarships and mentoring in Argentina.” *Journal of Research on Educational Effectiveness*, 13(2): 380–400.

Gertler, P. J., H. A. Patrinos, and M. Rubio-Codina. 2012. “Empowering parents to improve education: evidence from rural Mexico.” *Journal of Development Economics*, 99(1): 68–79.

Grissom, J. A., D. Kalogrides, and S. Loeb. 2015. “Using student test scores to measure principal performance.” *Educational Evaluation and Policy Analysis*, 37(1): 3–28.

Harris, D. 2005. “Comparison of 1-, 2-, and 3-parameter IRT models.” *Educational Measurement: Issues and Practice*, 8(1): 35–41.

IEA. 2015. “PIRLS 2016: Assessment framework.” TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

IEA. 2017. “TIMSS 2019: Assessment frameworks.” Edited by Mullis, I. V. S. & Martin, M. O. TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Jackson, C. K. 2020. “What do test scores miss? The importance of teacher effects on non-test-score outcomes.” *Journal of Political Economy.*

Kraft, M. A. 2019. “Teacher effects on complex cognitive skills and social-emotional competencies.” *Journal of Human Resources*, 54(1): 1–36.

Leaver, C., R. Lemos, and D. Scur. 2018. “Why does management matter? A theoretical framework.” *Unpublished manuscript.* Washington, DC: The World Bank.

Lemos, R., and D. Scur. 2016. “Developing management: An expanded evaluation tool for developing countries.” (RISE Working Paper No. 16/007). Washington, DC: Research on Improving Systems of Education (RISE).

Mbiti, I. M., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani. 2019. “Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania.” *The Quarterly Journal of Economics*, 134(3): 1627–1673.

Mbiti, Isaac M. 2016. “The need for accountability in education in developing countries.” *Journal of Economic Perspectives*, 30(3): 109–32.

Muralidharan, K., and A. Singh. 2018. “Understanding the flailing state: Experimental evidence from a large-scale school governance improvement program in India.” New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL) South Asia.

Muralidharan, K., and A. Singh. 2020. “Improving public sector management at scale: Experimental evidence on school governance in India.” *Unpublished manuscript.* San Diego, CA: University of California, San Diego.

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

Muralidharan, K., and V. Sundararaman. 2015. “The aggregate effect of school choice: Evidence from a two-stage experiment in India.” *The Quarterly Journal of Economics*, 130(3): 1011–1066.

Muralidharan, K., and Y. Zieleniak. 2014. "Chasing the syllabus: Measuring learning trajectories in developing countries with longitudinal data and item response theory." *Unpublished manuscript.* University of California, San Diego. San Diego, CA.

Muralidharan, K., A. Singh, and A. J. Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." *American Economic Review*, 109(4): 1–35.

National Education Law. 2006. "Nro. 26.206." Ciudad Autónoma de Buenos Aires, Argentina.

Pandey, P., S. Goyal, and V. Sundararaman. 2009. "Community participation in public schools: impact of information campaigns in three Indian states." *Education Economics*, 17(3): 355–375.

Pradhan, M., D. Suryadarma, A. Beatty, M. Wong, A. Gaduh, A. Alisjahbana, and R. P. Artha. 2014. "Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia." *American Economic Journal: Applied Economics*, 6(2): 105–26.

Pritchett, Lant, and Amanda Beatty. 2015. "Slow down, you're going too fast: Matching curricula to student skill levels." *International Journal of Educational Development*, 40: 276–288.

Romero, M., J. Sandefur, and W. Sandholtz. 2017. "Can outsourcing improve Liberia's schools? Preliminary results from year one of a three-year randomized evaluation of partnership schools for Liberia." (CGD Working Paper No. 462). Washington, DC: Center for Global Development (CGD).

SEE-MEDN. 2017. "Aprender 2016: Informe de resultados." Ciudad Autónoma de Buenos Aires: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

SEE-MEDN. 2018. "Aprender 2017: Informe de resultados, secundaria." Ciudad Autónoma de Buenos Aires: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

SEE-MEDN. 2019*a*. "Aprender 2018: Informe de resultados, Salta, 6to año primaria." Ciudad Autónoma de Buenos Aires: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

SEE-MEDN. 2019*b*. "Aprender 2018: Informe nacional de resultados, 6to año nivel primario." Ciudad Autónoma de Buenos Aires: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

Stata. 2017. "*Stata item response theory reference manual: Release 15* ." College Station, TX: StataCorp LLC.

UNESCO. 2020. "Global education monitoring report 2020. Inclusion and education: All means all." Paris, France: United Nations Educational, Scientific, and Cultural Organization (UNESCO).

UNESCO-LLECE. 2014. "Primera entrega de resultados: TERCE (Tercer Estudio Regional Comparativo y Explicativo)." Santiago, Chile: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), Oficina Regional de Educación para América Latina y el Caribe (OREALC).

Wong, H. L., R. Luo, L. Zhang, and S. Rozelle. 2013. "The impact of vouchers on preschool attendance and elementary school readiness: A randomized controlled trial in rural China." *Economics of Education Review*, 35: 53–65.

World Bank. 2018. *World Development Report 2018: Learning to realize education's promise.* Washington, DC: The World Bank.

Yen, W. M., and A. R. Fitzpatrick. 2006. "Item response theory." In Brennan, R. (Ed.) *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Table 1: Timeline of the study

(1)	(2)	(3)	(4)	(5)	(6)
		School participation rates			
		Control	Treatment		
		(Diagnostic	(Performance management)		
Month	Event	feedback)	Overall	Principals	Supervisors
A. 2014					
Apr	Students' performance in school (all primary grades)	100%	100%		
Oct	Independent assessments (grade 3)	96%	92%		
	Survey of students (grade 3)	94%	86%		
	Survey of teachers (grade 3)	94%	92%		
	Survey of principals	96%	82%		
B. 2015					
Mar	Delivery of school reports				
	Workshop 1: Using data from dashboards		88%	70%	88%
Apr	Students' performance in school (all primary grades)	100%	100%		
	Workshop 2: Developing school-improvement plans		80%	-	80%
May	Workshop 3: Implementing school-improvement plans		88%	76%	58%
Jun	Workshop 4: Conducting classroom observations		88%	78%	78%
Jul	Workshop 5: Tracking school-improvement plans		64%	-	64%
Aug	Workshop 6: Implementing effective teaching practices		74%	74%	-
Nov	Independent assessments (grades 3 and 4)	92%	92%		
	Survey of students (grades 3 and 4)	94%	92%		
	Survey of teachers (grades 3 and 4)	94%	94%		
	Survey of principals	88%	94%		
	Access to dashboards		78%		
C. 2016					
Apr	Delivery of school reports				
	Students' performance in school (all primary grades)	100%	100%		
	Workshop 7: Revising school-improvement plans		62%	58%	14%
May	Workshop 8: Implementing effective teaching practices		18%	18%	-
Jun	Workshop 9: Effective teaching practices in language		66%	56%	30%
Aug	Workshop 10: Effective teaching practices in math		64%	52%	26%
Aug	Workshop 11: Conducting classroom observations		78%	64%	42%
Nov	National assessment (grade 6)				
	Independent assessments (grades 3 and 5)	94%	92%		
	Survey of students (grades 3 and 5)	94%	90%		
	Survey of teachers (grades 3 and 5)	92%	84%		
	Survey of principals	96%	90%		
	Access to dashboards		8%		
D. 2017					
Apr	Students' performance in school (all primary grades)	100%	100%		
E. 2018					
Apr	Students' performance in school (all primary grades)	100%	100%		

Notes: (1) The table shows the timeline for the intervention and rounds of data collection for the study, including the month in which each event occurred (column 1), a brief description of the event (column 2), and the percentage of schools that participated in each event by experimental group (columns 3-6). (2) We display participation in the different elements of the performance-management intervention for either a principal or supervisor (column 4), and for principals (column 5) and supervisors (column 6) separately whenever appropriate. The dash (-) indicates that a group of individuals were not required to participate in a given event. (3) The school year in Argentina runs from February to November (see section 2.1).

Table 2: Balancing checks on students' performance in school, grades 3 to 5 (2014)

	(1) Control	(2) Treatment	(3) Difference
A. All primary-school grades			
Number of students enrolled	472.200	403.184	-64.783
	(355.483)	(289.141)	[39.970]
Percentage of students who passed the grade	96.861	96.934	0.085
	(3.528)	(3.634)	[0.709]
Percentage of students who failed the grade	2.637	2.317	-0.323
	(2.500)	(2.696)	[0.524]
Percentage of students who repeated the grade	3.630	2.503	-1.137
	(4.692)	(2.761)	[0.769]
Percentage of students who dropped out of school	0.502	0.749	0.238
	(1.558)	(2.710)	[0.433]
B. Grade 3			
Number of students enrolled	70.714	59.041	-10.164*
	(51.329)	(41.526)	[5.892]
Percentage of students who passed the grade	97.128	95.768	-1.364
	(4.193)	(6.394)	[1.089]
Percentage of students who failed the grade	2.688	3.679	0.992
	(3.868)	(5.637)	[0.977]
Percentage of students who repeated the grade	2.555	3.522	0.982
	(3.482)	(4.886)	[0.858]
Percentage of students who dropped out of school	0.185	0.554	0.372
	(1.024)	(3.607)	[0.531]
C. Grade 4			
Number of students enrolled	71.160	59.438	-12.037*
	(53.379)	(42.105)	[6.179]
Percentage of students who passed the grade	97.019	97.104	0.091
	(5.422)	(4.609)	[1.015]
Percentage of students who failed the grade	2.372	2.745	0.372
	(3.681)	(4.124)	[0.789]
Percentage of students who repeated the grade	3.407	2.699	-0.706
	(5.058)	(4.952)	[1.015]
Percentage of students who dropped out of school	0.608	0.151	-0.463
	(2.510)	(0.836)	[0.373]
D. Grade 5			
Number of students enrolled	67.240	57.776	-8.865
	(51.559)	(43.227)	[6.199]
Percentage of students who passed the grade	97.987	98.294	0.306
	(3.447)	(3.473)	[0.699]
Percentage of students who failed the grade	1.591	1.511	-0.074
	(2.285)	(2.841)	[0.515]
Percentage of students who repeated the grade	2.260	1.995	-0.259
	(4.267)	(3.376)	[0.774]
Percentage of students who dropped out of school	0.423	0.195	-0.233
	(2.830)	(0.887)	[0.422]
N (schools)	50	49	751

Notes: (1) The table shows the means and standard deviations of control schools (column 1) and treatment schools (column 2). It also tests for differences between these schools, using randomization fixed effects (column 3). Panel A shows results for all primary school students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) Dropout rates should be interpreted as a upper-bound estimate, as they actually refer to the percentage of students who leave their schools without asking for a pass to another school. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: ITT effect on students' performance in school, grades 3 to 5 (2015-2018)

	Percentage of students who passed the grade		Percentage of students who failed the grade		Percentage of students who repeated the grade		Percentage of students who dropped out of school	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2015								
Treatment	0.800	1.061*	-0.708	-0.952	0.217	0.191	-0.092	-0.098
	[0.722]	[0.611]	[0.679]	[0.577]	[0.723]	[0.649]	[0.131]	[0.116]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	97	97	97	97	96	96	97	97
Control mean	96.431		3.326		2.837		0.242	
B. 2016								
Treatment	0.707	0.956*	-0.852	-1.053**	-0.217	-0.168	0.145	0.143
	[0.618]	[0.525]	[0.603]	[0.530]	[0.804]	[0.800]	[0.172]	[0.173]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	96	96	96	96	97	96	96	96
Control mean	96.860		3.003		2.931		0.136	
C. 2017								
Treatment	2.162**	2.327***	-2.159**	-2.320***	-2.910**	-3.000**	-0.003	0.001
	[0.865]	[0.825]	[0.862]	[0.835]	[1.230]	[1.188]	[0.093]	[0.094]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	97	96	97	96	98	96	97	96
Control mean	95.748		4.150		5.626		0.101	
D. 2018								
Treatment	1.700*	1.741*	-1.442	-1.493*	-2.127**	-2.078**	-0.258	-0.234
	[0.995]	[0.979]	[0.881]	[0.874]	[0.894]	[0.870]	[0.220]	[0.217]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	98	99	98	98	97	99	98
Control mean	95.974		3.659		4.083		0.367	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in primary school. Panels A and B show these effects for 2015 and 2016 (the first and second year of the intervention, respectively) and Panels C and D for 2017 and 2018 (one and two years after the end of the intervention, respectively). Odd-numbered columns show estimates without any covariates other than randomization fixed effects. Even-numbered columns show estimates that also account for the pre-intervention school average for the corresponding indicator (e.g., column 3 shows results that account for passing rates in 2014). The control mean rows show the average value of the indicator for the control group. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: ITT effect on student achievement on independent assessments, grades 3 to 5 (2015-2016)

	Math						Reading					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Raw percent-correct score		Standardized percent-correct score		IRT-scaled score		Raw percent-correct score		Standardized percent-correct score		IRT-scaled score	
A. 2015												
Treatment	2.831	2.742	0.133	0.129	0.140	0.137	1.840	1.828	0.088	0.087	0.096	0.094
	[2.193]	[2.171]	[0.105]	[0.103]	[0.110]	[0.108]	[1.888]	[1.902]	[0.090]	[0.091]	[0.093]	[0.094]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	10529	10499	10529	10499	10529	10499	10528	10497	10528	10497	10528	10497
Control mean	67.412		-0.000		-0.000		68.402		-0.000		0.000	
B. 2016												
Treatment	0.384	0.131	0.020	0.006	0.023	0.006	-0.943	-0.538	-0.047	-0.023	-0.045	-0.018
	[1.959]	[1.929]	[0.104]	[0.102]	[0.108]	[0.107]	[1.522]	[1.452]	[0.090]	[0.085]	[0.090]	[0.085]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	10244	10220	10244	10220	10244	10220	10424	10404	10424	10404	10424	10404
Control mean	64.575		-0.000		0.000		68.112		-0.000		-0.000	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on independently administered student assessments of math (columns 1-6) and reading (columns 7-12), across grades 3 to 5 (see Table 1 for grades assessed each year. Panel A shows these effects for 2015 and Panel B for 2016. (2) Test scores are shown as raw percent-correct (columns 1-2 and 7-8), standardized percent correct (columns 3-4 and 9-10), and scaled scores from a two-parameter Item Response Theory (IRT) model (columns 5-6 and 11-12). (3) All standardization is with respect to the control group in the corresponding year. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: ITT effect on student achievement on national assessments, grade 6 (2016)

	Math	Reading
	(1)	(2)
	IRT-scaled score	IRT-scaled score
Treatment	0.013	0.011
	[0.101]	[0.108]
N (children)	4593	4485
Control mean	-0.152	-0.111

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on national student assessments of math and reading in grade 6 in 2016. (2) Test scores are shown as scaled scores from a two-parameter Item Response Theory (IRT) model. (3) All standardization is with respect to the average student in the country. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: Comparison between in- and out-of-sample schools on students' performance in school, grades 3 to 5 (2015-2018)

	(1) Percentage of students who passed the grade	(2) Percentage of students who failed the grade	(3) Percentage of students who repeated the grade	(4) Percentage of students who dropped out of school
A. 2015				
Diagnostic feedback	0.500	-0.008	0.017	-0.492**
	[0.632]	[0.547]	[0.441]	[0.202]
Performance management	1.319**	-0.728	0.237	-0.591***
	[0.615]	[0.572]	[0.660]	[0.178]
N (schools)	397	397	396	397
Out-of-sample mean	95.684	3.480	2.918	0.836
P-value (DF = PM)	0.277	0.297	0.759	0.524
B. 2016				
Diagnostic feedback	0.769	-0.251	0.120	-0.518***
	[0.578]	[0.529]	[0.593]	[0.150]
Performance management	1.464**	-1.095**	-0.104	-0.369*
	[0.568]	[0.529]	[0.653]	[0.205]
N (schools)	397	397	396	397
Out-of-sample mean	95.771	3.484	2.939	0.745
P-value (DF = PM)	0.292	0.178	0.780	0.410
C. 2017				
Diagnostic feedback	0.623	0.223	1.687	-0.846***
	[0.900]	[0.858]	[1.147]	[0.236]
Performance management	2.810***	-1.952***	-1.223**	-0.859***
	[0.593]	[0.481]	[0.608]	[0.242]
N (schools)	395	395	394	395
Out-of-sample mean	94.740	4.170	4.090	1.090
P-value (DF = PM)	0.017	0.013	0.017	0.933
D. 2018				
Diagnostic feedback	-0.055	0.798	0.740	-0.743**
	[0.932]	[0.764]	[0.798]	[0.335]
Performance management	1.663**	-0.650	-1.402**	-1.013***
	[0.709]	[0.597]	[0.561]	[0.270]
N (schools)	398	398	397	398
Out-of-sample mean	95.753	3.004	3.494	1.243
P-value (DF = PM)	0.090	0.099	0.016	0.306

Notes: (1) The table shows compares in- and out-of-sample schools on students' performance in school. Panels A and B show these effects for 2015 and 2016 (the two years of the interventions) and Panels C and D for 2017 and 2018 (the two years after the intervention). All estimations include geographic-area fixed effects. The out-of-sample mean show the average value of the indicator for public primary schools, which did not receive either intervention. The last row displays the p-value testing the null that the effects across both experimental groups are equal. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 7: Comparison between in- and out-of-sample schools on student achievement on national assessments, grade 6 (2016)

	Math	Reading
	(1) IRT-scaled score	(2) IRT-scaled score
Diagnostic feedback	0.052	0.067
	[0.089]	[0.092]
Performance management	0.064	0.078
	[0.053]	[0.062]
N (children)	16303	15981
Out-of-sample mean	-0.167	-0.165
P-value (DF = PM)	0.904	0.913

Notes: (1) The table shows the intent-to-treat (ITT) effect of the diagnostic-feedback and performance-management interventions on national student assessments of math and reading in grade 6 in 2016. (2) Test scores are shown as scaled scores from a two-parameter Item Response Theory (IRT) model. They are standardized with respect to the average student in the country. (3) The last row displays the p-value testing the null that the effects across both experimental groups are equal. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A Additional tables

Table A.1: Comparison between in- and out-of-sample schools on students' performance in school, grades 3 to 5 (2014)

	(1)	(2)	(3)	(4)	(5)	(6)
	All	Out-of-sample schools		In-sample	Col.(4)-	Col.(4)-
	schools	All	Non-rural	schools	Col.(2)	Col.(3)
A. All primary-school grades						
Number of students enrolled	213.973	179.951	359.934	438.040	258.089***	78.106**
	(288.370)	(266.744)	(301.771)	(324.553)	[29.663]	[35.566]
Percentage of students who passed the grade	95.304	95.061	95.854	96.897	1.837**	1.044
	(7.351)	(7.743)	(6.155)	(3.563)	[0.791]	[0.652]
Percentage of students who failed the grade	3.018	3.100	2.979	2.478	-0.622	-0.501
	(5.291)	(5.587)	(4.453)	(2.591)	[0.571]	[0.472]
Percentage of students who repeated the grade	3.347	3.389	3.043	3.072	-0.316	0.029
	(6.310)	(6.602)	(4.569)	(3.881)	[0.681]	[0.510]
Percentage of students who dropped out of school	1.678	1.839	1.167	0.624	-1.215**	-0.543
	(4.813)	(5.078)	(3.183)	(2.197)	[0.518]	[0.344]
B. Grade 3						
Number of students enrolled	33.724	28.593	52.422	64.878	36.284***	12.455**
	(42.355)	(39.314)	(43.250)	(46.813)	[4.410]	[5.130]
Percentage of students who passed the grade	95.295	95.105	95.277	96.461	1.356	1.184
	(10.638)	(11.254)	(8.677)	(5.400)	[1.158]	[0.931]
Percentage of students who failed the grade	3.547	3.608	3.572	3.173	-0.434	-0.398
	(8.720)	(9.201)	(6.559)	(4.817)	[0.950]	[0.719]
Percentage of students who repeated the grade	4.064	4.233	3.620	3.039	-1.194	-0.581
	(10.800)	(11.521)	(7.199)	(4.248)	[1.177]	[0.767]
Percentage of students who dropped out of school	1.158	1.287	1.151	0.366	-0.921	-0.785
	(5.911)	(6.278)	(4.958)	(2.621)	[0.643]	[0.523]
C. Grade 4						
Number of students enrolled	33.239	28.061	52.507	65.418	37.358***	12.912**
	(43.070)	(39.850)	(44.532)	(48.302)	[4.475]	[5.282]
Percentage of students who passed the grade	95.617	95.385	95.546	97.062	1.676	1.515*
	(11.669)	(12.395)	(8.159)	(5.010)	[1.263]	[0.870]
Percentage of students who failed the grade	3.427	3.567	3.561	2.557	-1.010	-1.004
	(10.150)	(10.817)	(6.446)	(3.891)	[1.099]	[0.686]
Percentage of students who repeated the grade	3.362	3.411	3.296	3.060	-0.351	-0.236
	(9.882)	(10.460)	(7.032)	(4.993)	[1.076]	[0.766]
Percentage of students who dropped out of school	0.956	1.048	0.893	0.382	-0.666	-0.511
	(5.401)	(5.764)	(3.276)	(1.883)	[0.585]	[0.347]
D. Grade 5						
Number of students enrolled	32.098	27.211	51.469	62.556	35.345***	11.087**
	(42.340)	(39.334)	(44.283)	(47.613)	[4.392]	[5.219]
Percentage of students who passed the grade	96.216	95.904	95.959	98.139	2.235*	2.180**
	(11.417)	(12.205)	(8.563)	(3.446)	[1.235]	[0.884]
Percentage of students who failed the grade	2.175	2.277	2.634	1.551	-0.725	-1.082*
	(7.223)	(7.716)	(6.093)	(2.562)	[0.783]	[0.630]
Percentage of students who repeated the grade	2.601	2.677	2.747	2.129	-0.548	-0.618
	(8.741)	(9.290)	(6.285)	(3.835)	[0.947]	[0.669]
Percentage of students who dropped out of school	1.609	1.820	1.408	0.310	-1.510	-1.098**
	(8.715)	(9.342)	(5.358)	(2.098)	[0.943]	[0.552]
N (schools)	751	652	305	99	751	404

Notes: (1) The table shows the means and standard deviations of all public primary schools in Salta (column 1), non-RCT schools (columns 2-3), and RCT schools (column 4). It also tests for differences between all non-RCT and RCT schools (column 5), and between non-rural, non-RCT schools and RCT schools (column 6). Panel A shows results for all primary school students, Panel B for grade 3 students, and Panel C for grade 5 students. (2) Dropout rates should be interpreted as a upper-bound estimate, as they actually refer to the percentage of students who leave their schools without asking for a pass to another school. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2: Balancing checks on independent assessments, grade 3 (2014)

	(1) Control	(2) Treatment	(3) Difference
A. Math			
Raw percent-correct score	61.175	62.863	1.564
	(21.577)	(20.969)	[2.229]
Standardized percent-correct score	0.000	0.078	0.073
	(1.000)	(0.972)	[0.103]
Standardized IRT-scaled score	0.000	0.092	0.087
	(1.000)	(0.981)	[0.105]
N (students)	6,530	5,617	12,147
B. Reading			
Raw percent-correct score	63.003	62.375	-0.822
	(21.603)	(21.023)	[1.946]
Standardized percent-correct score	0.000	-0.029	-0.038
	(1.000)	(0.973)	[0.090]
Standardized IRT-scaled score	0.000	-0.024	-0.032
	(1.000)	(0.966)	[0.090]
N (students)	6,572	5,523	12,095

Notes: (1) The table shows the means and standard deviations of grade 3 students in control schools (column 1) and treatment schools (column 2). It also tests for differences between these schools, using randomization fixed effects (column 3). (2) Test scores are shown as raw percent-correct, standardized percent correct, and scaled scores from a two-parameter Item Response Theory (IRT) model. Scores are standardized with respect to the control group on each year. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Balancing checks on student survey, grade 3 (2014)

	(1) Control	(2) Treatment	(3) Difference
Female	0.488	0.481	-0.008
	(0.500)	(0.500)	[0.017]
Overage	0.485	0.467	-0.016
	(0.500)	(0.499)	[0.021]
Uses computer at school	0.199	0.200	0.006
	(0.399)	(0.400)	[0.050]
Has access to Internet at home	0.469	0.464	-0.012
	(0.499)	(0.499)	[0.042]
Uses computer at home	0.644	0.634	-0.016
	(0.479)	(0.482)	[0.033]
N (students)	2,863	2,405	5,268

Notes: (1) The table shows the means and standard deviations of grade 3 students in control schools (column 1) and treatment schools (column 2). It also tests for differences between these students, using randomization fixed effects (column 3). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Balancing checks on teacher survey, grade 3 (2014)

	(1) Control	(2) Treatment	(3) Difference
Age	43.024	42.342	-0.692
	(7.973)	(7.560)	[0.967]
Teaches Spanish	0.916	0.974	0.059*
	(0.278)	(0.160)	[0.034]
Teaches math	0.901	0.957	0.058
	(0.300)	(0.205)	[0.041]
Has a university degree	0.029	0.043	0.013
	(0.168)	(0.204)	[0.022]
Has 5+ years of teaching experience	0.790	0.752	-0.039
	(0.409)	(0.434)	[0.054]
Has 5+ years of teaching experience at this school	0.475	0.397	-0.079
	(0.501)	(0.491)	[0.069]
Teaches at another school	0.345	0.250	-0.096*
	(0.477)	(0.435)	[0.055]
Has another non-teaching job	0.167	0.120	-0.047
	(0.374)	(0.326)	[0.043]
N (teachers)	141	120	261

Notes: (1) The table shows the means and standard deviations of grade 3 teachers in control schools (column 1) and treatment schools (column 2). It also tests for differences between these teachers, using randomization fixed effects (column 3). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Balancing checks on principal survey (2014)

	(1) Control	(2) Treatment	(3) Difference
Female	0.917	0.805	-0.109
	(0.279)	(0.401)	[0.071]
Age	51.417	49.974	-1.446
	(5.511)	(6.583)	[1.323]
Has 5+ years of experience	0.479	0.561	0.080
	(0.505)	(0.502)	[0.106]
Has 5+ years of experience at this school	0.184	0.244	0.060
	(0.391)	(0.435)	[0.086]
Teaches at another school	0.306	0.341	0.035
	(0.466)	(0.480)	[0.097]
Has another non-school related job	0.204	0.073	-0.131*
	(0.407)	(0.264)	[0.071]
Is formally appointed	0.245	0.317	0.072
	(0.434)	(0.471)	[0.097]
Has a university degree	0.204	0.075	-0.128*
	(0.407)	(0.267)	[0.071]
N (principals)	49	41	90

Notes: (1) The table shows the means and standard deviations of principals in control schools (column 1) and treatment schools (column 2). It also tests for differences between these principals, using randomization fixed effects (column 3). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: Implementation fidelity based on principal survey (2016)

	(1)	(2)	(3)	(4)	(5)	(6)
	I received a report in 2015	I received a report in 2016	I developed a school improvement plan	I set goals based on that plan	I attended workshops at least once	I used the dashboard at least once
Treatment	0.328***	0.427***	0.044	0.046	0.157*	0.039
	[0.104]	[0.100]	[0.031]	[0.044]	[0.087]	[0.102]
N (principals)	71	76	88	88	93	93
Control mean	0.545	0.410	0.956	0.932	0.688	0.604

Notes: (1) The table compares self-reported implementation fidelity from surveys of principals. The control mean displays the share of principals in the control group who indicated engaging in a given behavior and the coefficient on the treatment dummy indicates the additional share of principals in the treatment group that engaged in that behavior. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: Participation in school-management programs based on principal survey (2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Nueva Escuela	Matemática para Todos	Programa Integral por la Igualdad Educativa (PIIE)	Proyecto de Mejoramiento de la Educación Rural (PROMER)	Programa de Mejoramiento de la Equidad y Calidad de la Educación (PROMEDU)	Gestión Escolar para la Mejora de los Aprendizajes (GEMA)	Primaria Digital
Treatment	0.063	-0.012	0.058	-0.012	0.046	-0.078	0.006
	[0.048]	[0.090]	[0.080]	[0.087]	[0.071]	[0.115]	[0.106]
N (principals)	88	77	81	68	66	76	79
Control mean	0.913	0.500	0.500	0.132	0.056	0.628	0.682

Notes: (1) The table compares self-reported participation in other school-management initiatives surveys of principals. The control mean displays the share of principals in the control group who indicated their school participated in a given initiative and the coefficient on the treatment dummy indicates the additional share of principals in the treatment group that participated in that initiative. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: ITT effect on students' performance in school, grade 3 (2015-2018)

	Percentage of students who passed the grade		Percentage of students who failed the grade		Percentage of students who repeated the grade		Percentage of students who dropped out of school	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2015								
Treatment	1.242	1.418	-1.060	-1.232	0.374	-0.380	-0.182*	-0.190*
	[0.911]	[0.876]	[0.917]	[0.876]	[1.235]	[1.083]	[0.103]	[0.105]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	97	97	97	97	98	97	97	97
Control mean	96.113		3.666		3.566		0.220	
B. 2016								
Treatment	0.081	0.424	-0.174	-0.398	0.135	-0.092	0.092	0.102
	[0.972]	[0.934]	[0.965]	[0.927]	[0.897]	[0.909]	[0.269]	[0.284]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	98	97	98	97	99	98	98	97
Control mean	96.220		3.538		2.854		0.243	
C. 2017								
Treatment	1.774	2.267	-1.671	-2.041	-2.302*	-2.540*	-0.103	-0.090
	[1.777]	[1.773]	[1.742]	[1.755]	[1.364]	[1.326]	[0.264]	[0.270]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	98	99	98	98	97	99	98
Control mean	94.230		5.486		5.382		0.285	
D. 2018								
Treatment	0.048	0.225	0.118	-0.046	-2.179**	-2.383**	-0.166	-0.182
	[1.240]	[1.111]	[1.189]	[1.057]	[1.081]	[1.046]	[0.203]	[0.211]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	98	99	98	98	98	99	98
Control mean	96.822		2.876		4.054		0.302	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in grade 3. Panels A and B show these effects for 2015 and 2016 (the first and second year of the intervention, respectively) and Panels C and D for 2017 and 2018 (one and two years after the end of the intervention, respectively). Odd-numbered columns show estimates without any covariates other than randomization fixed effects. Even-numbered columns show estimates that also account for the prior-year school average for the corresponding indicator (e.g., column 3 shows results that account for passing rates in 2014). The control mean rows show the average value of the indicator for the control group. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: ITT effect on students' performance in school, grade 4 (2015-2018)

	Percentage of students who passed the grade		Percentage of students who failed the grade		Percentage of students who repeated the grade		Percentage of students who dropped out of school	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2015								
Treatment	0.718	0.676	-0.610	-0.771	-0.834	-0.623	-0.108	0.086
	[0.827]	[0.679]	[0.733]	[0.655]	[0.591]	[0.552]	[0.208]	[0.122]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	99	99	99	98	97	99	99
Control mean	96.967		2.809		2.749		0.224	
B. 2016								
Treatment	1.079	1.060	-1.419**	-1.482**	-0.118	-0.057	0.339	0.343
	[0.720]	[0.694]	[0.688]	[0.667]	[1.228]	[1.243]	[0.243]	[0.237]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	99	99	99	98	97	99	99
Control mean	96.885		3.050		2.907		0.065	
C. 2017								
Treatment	1.572*	1.569*	-1.608*	-1.676**	-3.523**	-3.315**	0.036	0.036
	[0.842]	[0.836]	[0.841]	[0.822]	[1.360]	[1.298]	[0.037]	[0.038]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	98	98	98	98	99	98	98	98
Control mean	96.770		3.216		5.599		0.014	
D. 2018								
Treatment	2.103	2.086	-2.114	-2.169*	-1.583	-1.479	0.011	0.029
	[1.289]	[1.284]	[1.276]	[1.269]	[1.402]	[1.425]	[0.064]	[0.070]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	99	99	99	99	98	99	99
Control mean	95.649		4.294		4.006		0.057	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in school in grade 4. Panel A shows these effects for 2015 (the first year of the intervention), Panel B for 2016 (the second year of the intervention), and Panel C for 2017 (one year after the end of the intervention). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.10: ITT effect on students' performance in school, grade 5 (2015-2018)

	Percentage of students who passed the grade		Percentage of students who failed the grade		Percentage of students who repeated the grade		Percentage of students who dropped out of school	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. 2015								
Treatment	0.416	0.319	-0.431	-0.388	1.019	1.072	0.015	0.022
	[1.024]	[1.006]	[0.971]	[0.928]	[0.788]	[0.766]	[0.196]	[0.199]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	99	99	99	98	98	99	99
Control mean	96.349		3.378		2.081		0.273	
B. 2016								
Treatment	1.021	1.061	-1.003	-0.973	-0.658	-0.535	-0.018	-0.015
	[0.757]	[0.720]	[0.751]	[0.726]	[1.371]	[1.376]	[0.105]	[0.107]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	97	97	97	97	98	98	97	97
Control mean	97.617		2.288		2.913		0.095	
C. 2017								
Treatment	2.166**	2.169**	-2.225**	-2.221**	-3.048*	-2.717*	0.059	0.037
	[0.893]	[0.896]	[0.891]	[0.897]	[1.700]	[1.395]	[0.042]	[0.028]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	98	98	98	98	99	99	98	98
Control mean	96.429		3.571		5.898		0.000	
D. 2018								
Treatment	2.947**	2.866**	-2.328***	-2.294***	-2.430***	-2.315***	-0.619	-0.610
	[1.123]	[1.126]	[0.869]	[0.839]	[0.864]	[0.802]	[0.562]	[0.570]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
N (schools)	99	99	99	99	99	99	99	99
Control mean	95.452		3.807		4.024		0.740	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in school in grade 5. Panel A shows these effects for 2015 (the first year of the intervention), Panel B for 2016 (the second year of the intervention), and Panel C for 2017 (one year after the end of the intervention). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.11: ITT effect on student achievement on independent assessments, grade 3 (2015-2016)

	Math						Reading					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Raw percent-correct score		Standardized percent-correct score		IRT-scaled score		Raw percent-correct score		Standardized percent-correct score		IRT-scaled score	
A. 2015												
Treatment	4.009	3.927	0.186	0.182	0.206*	0.204*	1.982	2.003	0.093	0.094	0.114	0.114
	[2.421]	[2.366]	[0.112]	[0.110]	[0.119]	[0.115]	[1.995]	[2.006]	[0.093]	[0.094]	[0.094]	[0.094]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	5000	4987	5000	4987	5000	4987	4999	4985	4999	4985	4999	4985
Control mean	68.206		0.000		-0.000		70.284		-0.000		0.000	
B. 2016												
Treatment	1.100	0.976	0.053	0.047	0.059	0.053	-1.284	-0.932	-0.069	-0.050	-0.066	-0.044
	[2.121]	[2.119]	[0.102]	[0.102]	[0.104]	[0.104]	[1.870]	[1.801]	[0.100]	[0.097]	[0.101]	[0.097]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	4965	4956	4965	4956	4965	4956	5099	5091	5099	5091	5099	5091
Control mean	68.453		0.000		-0.000		70.655		0.000		-0.000	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on independently administered student assessments of math (columns 1-3) and reading (columns 4-6), in grade 3 (see Table 1 for grades assessed each year. Panel A shows these effects for 2015 and Panel B for 2016. (2) Test scores are shown as raw percent-correct (columns 1 and 4), standardized percent correct (columns 2 and 5), and scaled scores from a two-parameter Item Response Theory (IRT) model. (3) Test in columns 2-3 and 5-6 are standardized with respect to the control group in the corresponding year. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.12: ITT effect on student achievement on independent assessments, grade 4 (2015)

	Math						Reading					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Raw percent-correct score		Standardized percent-correct score		IRT-scaled score		Raw percent-correct score		Standardized percent-correct score		IRT-scaled score	
Treatment	1.752	1.654	0.086	0.081	0.080	0.075	1.723	1.699	0.084	0.083	0.079	0.077
	[2.396]	[2.401]	[0.117]	[0.117]	[0.122]	[0.122]	[2.397]	[2.406]	[0.117]	[0.117]	[0.122]	[0.122]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	5529	5512	5529	5512	5529	5512	5529	5512	5529	5512	5529	5512
Control mean	66.697		-0.000		-0.000		66.695		-0.000		-0.000	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on independently administered student assessments of math (columns 1-3) and reading (columns 4-6), in grade 4 (see Table 1 for grades assessed each year. (2) Test scores are shown as raw percent-correct (columns 1 and 4), standardized percent correct (columns 2 and 5), and scaled scores from a two-parameter Item Response Theory (IRT) model. (3) Test in columns 2-3 and 5-6 are standardized with respect to the control group in the corresponding year. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.13: ITT effect on student achievement on independent assessments, grade 5 (2016)

	Math						Reading					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Raw percent-correct score		Standardized percent-correct score		IRT-scaled score		Raw percent-correct score		Standardized percent-correct score		IRT-scaled score	
Treatment	-0.173	-0.610	-0.010	-0.036	-0.010	-0.042	-0.421	0.037	-0.028	0.002	-0.026	0.007
	[2.128]	[2.085]	[0.124]	[0.121]	[0.130]	[0.128]	[1.428]	[1.365]	[0.094]	[0.090]	[0.093]	[0.088]
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N (children)	5279	5264	5279	5264	5279	5264	5325	5313	5325	5313	5325	5313
Control mean	60.887		-0.000		0.000		65.582		-0.000		-0.000	

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on independently administered student assessments of math (columns 1-3) and reading (columns 4-6), in grade 5 (see Table 1 for grades assessed each year. (2) Test scores are shown as raw percent-correct (columns 1 and 4), standardized percent correct (columns 2 and 5), and scaled scores from a two-parameter Item Response Theory (IRT) model. (3) Test in columns 2-3 and 5-6 are standardized with respect to the control group in the corresponding year. (4) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.14: ITT effect on student-reported teacher attendance, punctuality, and time allocation in math, grades 3 to 5 (2015-2016)

	(1) Teacher arrived late to class	(2) Teacher was absent to school	(3) Teacher started lesson late	(4) Teacher ended lesson early	(5) Teacher left school early
A. 2015					
Treatment	0.000	0.006	0.007	0.023	0.007
	[0.030]	[0.034]	[0.030]	[0.031]	[0.030]
N (children)	7688	8013	7567	7530	7480
Control mean	0.250	0.488	0.374	0.390	0.310
B. 2016					
Treatment	0.004	-0.011	-0.015	0.011	0.023
	[0.020]	[0.030]	[0.018]	[0.018]	[0.019]
N (children)	8223	8240	8225	8178	8298
Control mean	0.192	0.481	0.255	0.296	0.225

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on teacher effort as indicated by the student surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Students were asked to indicate how frequently their teachers engaged in a number of activities. The results above show the share of students who reported their teacher engaged in these behaviors two or more times in the two weeks prior to the survey. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: ITT effect on student-reported teaching activities in math, grades 3 to 5 (2015-2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	I used a textbook	Teacher assigned homework	I copied from the blackboard	I solved problems	I worked in groups	I practiced exercises on the blackboard	Teacher explained topics	Teacher assigned practice tests	Teacher graded homework
A. 2015									
Treatment	-0.006	0.011	-0.001	-0.004	0.016	0.008	0.003	0.013	-0.001
	[0.028]	[0.014]	[0.015]	[0.008]	[0.021]	[0.013]	[0.011]	[0.015]	[0.008]
N (children)	8440	8586	8516	8549	8407	8336	8379	8338	8566
Control mean	0.703	0.877	0.860	0.916	0.789	0.827	0.906	0.853	0.923
B. 2016									
Treatment	0.024	-0.014	0.002	0.001	0.022	0.006	-0.007	0.001	-0.005
	[0.032]	[0.021]	[0.012]	[0.010]	[0.022]	[0.012]	[0.009]	[0.014]	[0.008]
N (children)	8285	8346	8205	8243	8086	8193	8090	8195	8351
Control mean	0.651	0.875	0.863	0.913	0.771	0.830	0.920	0.844	0.925

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on teacher effort as indicated by the student surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Students were asked to indicate how frequently their math teacher engaged in a number of activities. The results above show the share of students who reported their math teacher engaged in these behaviors two or more times in the two weeks prior to the survey. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.16: ITT effect on student-reported teaching activities in Spanish, grades 3 to 5 (2015-2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	I used a textbook	Teacher assigned homework	I copied from the blackboard	I wrote something	I worked in groups	Teacher dictated texts	Teacher explained topics	Teacher assigned practice tests	Teacher graded homework
A. 2015									
Treatment	-0.005	0.003	0.017	0.013	0.001	-0.011	0.002	0.007	-0.002
	[0.022]	[0.015]	[0.017]	[0.019]	[0.015]	[0.023]	[0.012]	[0.015]	[0.008]
N (children)	8788	8959	8776	8733	8779	8621	8653	8671	8880
Control mean	0.787	0.862	0.841	0.773	0.842	0.820	0.902	0.853	0.925
B. 2016									
Treatment	0.006	-0.013	-0.021*	-0.022	0.005	-0.008	-0.009	-0.009	0.004
	[0.017]	[0.024]	[0.012]	[0.019]	[0.018]	[0.016]	[0.008]	[0.007]	[0.013]
N (children)	8438	8465	8310	8381	8325	8345	8278	8470	8335
Control mean	0.769	0.839	0.873	0.776	0.842	0.841	0.920	0.938	0.853

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on teacher effort as indicated by the student surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Students were asked to indicate how frequently their Spanish teacher engaged in a number of activities. The results above show the share of students who reported their Spanish teacher engaged in these behaviors two or more times in the two weeks prior to the survey. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.17: ITT effect on student ratings of teacher effectiveness, grades 3 to 5 (2015-2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Demonstrating interest in students	Managing classroom discipline	Clarifying difficult concepts or tasks	Pushing students to do their best	Delivering engaging lessons	Engaging students in discussions	Summarizing material at the end of lessons
A. 2015							
Treatment	-0.073	-0.049	-0.021	-0.025	-0.045	-0.000	-0.057
	[0.055]	[0.071]	[0.059]	[0.068]	[0.050]	[0.067]	[0.068]
N (children)	9595	9541	9557	9521	9482	9426	9363
Control mean	0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000
B. 2016							
Treatment	0.010	0.029	0.019	-0.002	0.015	0.004	-0.043
	[0.044]	[0.057]	[0.050]	[0.047]	[0.046]	[0.054]	[0.050]
N (children)	8954	8932	8914	8890	8858	8834	8827
Control mean	0.000	-0.000	-0.000	0.000	-0.000	0.000	0.000

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on teacher effort as indicated by the student surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Students were asked to indicate how frequently their teacher engaged in a number of activities (e.g., "my teacher gives me enough time to explain my answers"). The results above show the mean standardized score for each group of behaviors. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.18: ITT effect on principal-reported use of assessment results (2015-2016)

	(1)	(2)	(3)	(4)	(5)
	I am evaluated based on assessment results	Teachers are evaluated based on assesssment results	I make curriculum changes based on assessment results	I share assessment results with parents	I make assessment results public
A. 2015					
Treatment	0.101	0.166	0.112*	0.301***	0.298***
	[0.112]	[0.110]	[0.062]	[0.088]	[0.104]
N (principals)	83	84	82	83	81
Control mean	0.486	0.421	0.865	0.632	0.378
B. 2016					
Treatment	0.032	0.034	0.026	0.336***	0.228**
	[0.110]	[0.111]	[0.040]	[0.082]	[0.105]
N (principals)	85	85	86	86	85
Control mean	0.500	0.477	0.952	0.619	0.452

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on teacher effort as indicated by the student surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Students were asked to indicate how frequently their teacher engaged in a number of activities (e.g., "my teacher gives me enough time to explain my answers"). The results above show the mean standardized score for each group of behaviors. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.19: ITT effect on principal-reported instructional leadership (2015-2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	I observed a full lesson	I observed part of a lesson	I gave a teacher feeedback	I led a demonstration lesson	I acknowledged a peer's work	I led a training at my school	I supported collaboration among teachers	I held teachers accountable for learning
A. 2015								
Treatment	0.002	-0.052	0.036	0.029	-0.036	-0.059	0.031	-0.014
	[0.103]	[0.086]	[0.087]	[0.105]	[0.072]	[0.105]	[0.074]	[0.068]
N (principals)	91	91	91	91	91	91	91	91
Control mean	0.636	0.818	0.773	0.545	0.886	0.591	0.841	0.864
B. 2016								
Treatment	0.022	-0.037	0.005	-0.050	-0.032	-0.054	-0.011	-0.008
	[0.099]	[0.084]	[0.087]	[0.105]	[0.066]	[0.105]	[0.069]	[0.058]
N (principals)	93	93	93	93	93	93	93	93
Control mean	0.667	0.812	0.771	0.583	0.896	0.542	0.875	0.917

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on instructional leadership as indicated by the principal surveys. Panel A shows these effects for 2015 and Panel B for 2016. (2) Principals were asked to indicate how frequently their teacher engaged in a number of activities. The results above show the share of principals who reported engaging in these behaviors two or more times in the month prior to the survey. (3) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.20: Comparison between in- and out-of-sample schools on students' performance in school, grade 3 (2015-2018)

	(1) Percentage of students who passed the grade	(2) Percentage of students who failed the grade	(3) Percentage of students who repeated the grade	(4) Percentage of students who dropped out of school
A. 2015				
Diagnostic feedback	0.922	-0.409	0.092	-0.514**
	[0.807]	[0.772]	[0.723]	[0.224]
Performance management	2.188***	-1.485**	0.469	-0.702***
	[0.769]	[0.729]	[1.111]	[0.199]
N (schools)	398	398	399	398
Out-of-sample mean	94.945	4.250	3.608	0.805
P-value (DF = PM)	0.182	0.247	0.758	0.158
B. 2016				
Diagnostic feedback	0.689	-0.298	-0.736	-0.391*
	[0.797]	[0.759]	[0.676]	[0.212]
Performance management	0.763	-0.466	-0.621	-0.297
	[0.818]	[0.804]	[0.810]	[0.272]
N (schools)	399	399	401	399
Out-of-sample mean	95.325	3.964	3.687	0.711
P-value (DF = PM)	0.939	0.861	0.900	0.726
C. 2017				
Diagnostic feedback	0.042	0.780	0.819	-0.822**
	[1.523]	[1.477]	[1.282]	[0.368]
Performance management	1.836	-0.901	-1.483*	-0.935**
	[1.210]	[1.104]	[0.773]	[0.370]
N (schools)	399	399	397	399
Out-of-sample mean	93.841	4.876	4.676	1.283
P-value (DF = PM)	0.314	0.332	0.089	0.694
D. 2018				
Diagnostic feedback	1.406	-0.624	0.413	-0.781**
	[0.997]	[0.877]	[1.023]	[0.328]
Performance management	1.474	-0.516	-1.804***	-0.958***
	[1.072]	[1.015]	[0.648]	[0.291]
N (schools)	400	400	399	400
Out-of-sample mean	95.083	3.694	3.858	1.223
P-value (DF = PM)	0.957	0.927	0.041	0.458

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in grade 3. Panels A and B show these effects for 2015 and 2016 (the first and second year of the intervention, respectively) and Panels C and D for 2017 and 2018 (one and two years after the end of the intervention, respectively). The control mean rows show the average value of the indicator for the control group. (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.21: Comparison between in- and out-of-sample schools on students' performance in school, grade 4 (2015-2018)

	(1)	(2)	(3)	(4)
	Percentage of students who passed the grade	Percentage of students who failed the grade	Percentage of students who repeated the grade	Percentage of students who dropped out of school
A. 2015				
Diagnostic feedback	0.946	-0.512	-0.175	-0.434*
	[0.818]	[0.691]	[0.524]	[0.254]
Performance management	1.690**	-1.143*	-1.044**	-0.547***
	[0.706]	[0.668]	[0.515]	[0.197]
N (schools)	402	402	402	402
Out-of-sample mean	95.818	3.433	3.003	0.750
P-value (DF = PM)	0.396	0.411	0.154	0.599
B. 2016				
Diagnostic feedback	0.853	-0.300	0.289	-0.553***
	[0.733]	[0.697]	[0.790]	[0.177]
Performance management	1.954***	-1.738***	0.171	-0.216
	[0.708]	[0.645]	[1.042]	[0.294]
N (schools)	403	403	400	403
Out-of-sample mean	95.700	3.616	2.721	0.684
P-value (DF = PM)	0.149	0.046	0.922	0.169
C. 2017				
Diagnostic feedback	1.767*	-0.841	1.470	-0.926***
	[0.903]	[0.853]	[1.313]	[0.286]
Performance management	3.326***	-2.441***	-2.061***	-0.885***
	[0.706]	[0.596]	[0.647]	[0.288]
N (schools)	397	397	398	397
Out-of-sample mean	94.572	4.369	4.316	1.059
P-value (DF = PM)	0.083	0.063	0.009	0.747
D. 2018				
Diagnostic feedback	-0.471	1.669	0.630	-1.198***
	[1.108]	[1.048]	[0.986]	[0.311]
Performance management	1.648*	-0.447	-0.949	-1.201***
	[0.941]	[0.853]	[1.176]	[0.312]
N (schools)	400	400	401	400
Out-of-sample mean	95.847	2.727	3.513	1.426
P-value (DF = PM)	0.102	0.095	0.257	0.985

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in school in grade 4. Panel A shows these effects for 2015 (the first year of the intervention), Panel B for 2016 (the second year of the intervention), and Panel C for 2017 (one year after the end of the intervention). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.22: Comparison between in- and out-of-sample schools on students' performance in school, grade 5 (2015-2018)

	(1) Percentage of students who passed the grade	(2) Percentage of students who failed the grade	(3) Percentage of students who repeated the grade	(4) Percentage of students who dropped out of school
A. 2015				
Diagnostic feedback	-0.176	0.715	0.053	-0.539**
	[0.806]	[0.737]	[0.463]	[0.249]
Performance management	0.251	0.282	1.070	-0.533**
	[0.870]	[0.813]	[0.749]	[0.246]
N (schools)	402	402	401	402
Out-of-sample mean	96.274	2.794	2.071	0.932
P-value (DF = PM)	0.677	0.653	0.193	0.977
B. 2016				
Diagnostic feedback	0.988	-0.371	0.661	-0.617***
	[0.793]	[0.752]	[1.140]	[0.201]
Performance management	2.036***	-1.392**	-0.005	-0.644***
	[0.634]	[0.565]	[0.882]	[0.204]
N (schools)	401	401	401	401
Out-of-sample mean	96.260	2.916	2.424	0.824
P-value (DF = PM)	0.200	0.187	0.621	0.860
C. 2017				
Diagnostic feedback	0.283	0.525	2.588*	-0.808***
	[0.899]	[0.860]	[1.493]	[0.232]
Performance management	2.509***	-1.742***	-0.458	-0.767***
	[0.643]	[0.574]	[0.967]	[0.243]
N (schools)	399	399	398	399
Out-of-sample mean	95.825	3.268	3.513	0.907
P-value (DF = PM)	0.017	0.013	0.071	0.675
D. 2018				
Diagnostic feedback	-0.988	1.412*	1.063	-0.424
	[1.111]	[0.818]	[0.828]	[0.625]
Performance management	1.980***	-0.923	-1.373***	-1.057***
	[0.711]	[0.560]	[0.525]	[0.307]
N (schools)	400	400	400	400
Out-of-sample mean	96.169	2.532	3.031	1.299
P-value (DF = PM)	0.010	0.007	0.005	0.279

Notes: (1) The table shows the intent-to-treat (ITT) effect of the performance-management intervention on administrative data on students' performance in school in grade 5. Panel A shows these effects for 2015 (the first year of the intervention), Panel B for 2016 (the second year of the intervention), and Panel C for 2017 (one year after the end of the intervention). (2) * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix B Additional information on instruments

B.1 Student assessments

B.1.1 Design

The student assessments that we used in this study were developed by a domestic think tank (the *Centro de Estudios en Políticas Públicas*). The think tank created its own items and drew on publicly-released items from the national student assessment (called the *Operativo Nacional de Evaluación* at the time of the study) (see, for example, DiNIECE 2009, 2012).

The content and skills evaluated in the assessments in the present study are similar to those in national and international assessments. As we discuss in section 3, our assessments were based on both national and provincial standards (the *Contenidos Básicos Comunes*, *Núcleos de Aprendizaje Prioritarios*, and *Diseño Curricular de La Rioja*). Additionally, the distribution of items across content and cognitive domains in our assessments are consistent with those of the current national assessment (*Aprender*) and with international assessments of primary school students such as the Trends in International Math and Science Study (TIMSS) and the Progress in International Reading Study (PIRLS) (see Table B.1).

B.1.2 Scaling and linking

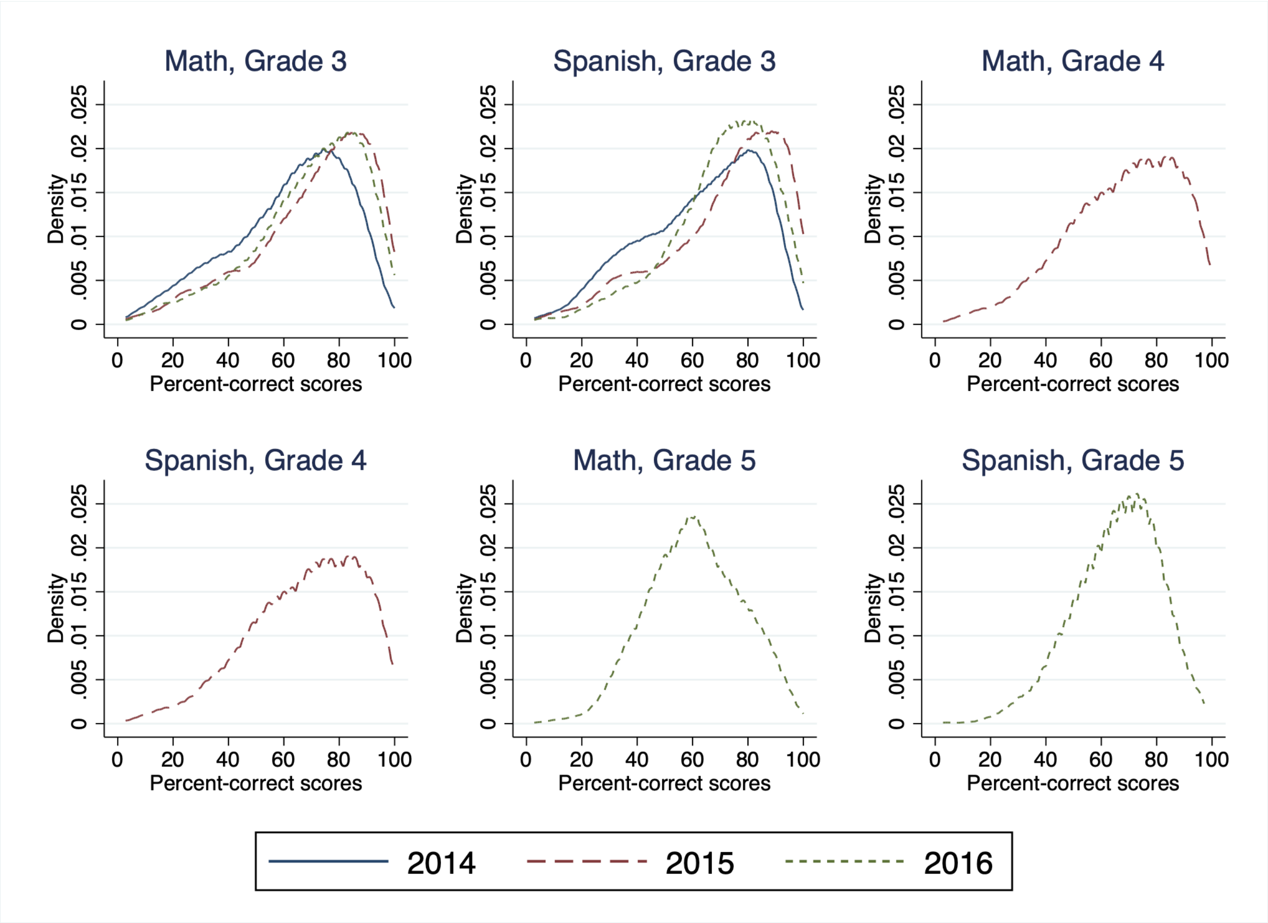
The think tank that designed the assessments scored all items dichotomously. We used a two-parameter logistic (2PL) Item Response Theory (IRT) model to scale and link the results (Harris 2005).[32] This model allows us to account for differences between items (specifically, differences in their difficulty and capacity to distinguish between students of similar ability). It also allows us to capitalize on common items across data collection rounds (within subjects and grades) to map assessment results for all three years of the study (2013 to 2015) onto the same scale. We standardized all IRT-scaled scores to have a mean of 0 and a standard deviation of 1 with respect to the control group in 2014.

B.1.3 Distributions of percent-correct and IRT-scaled scores

The design, scaling, and linking processes described above were successful in producing well-behaved distributions in all grades, subjects, and years of the study, with little evidence of "floor" effects (i.e., a high concentration of students with no correct answers) or "ceiling" effects (i.e., a high concentration of students with perfect scores) (Figures **??** and **??**).

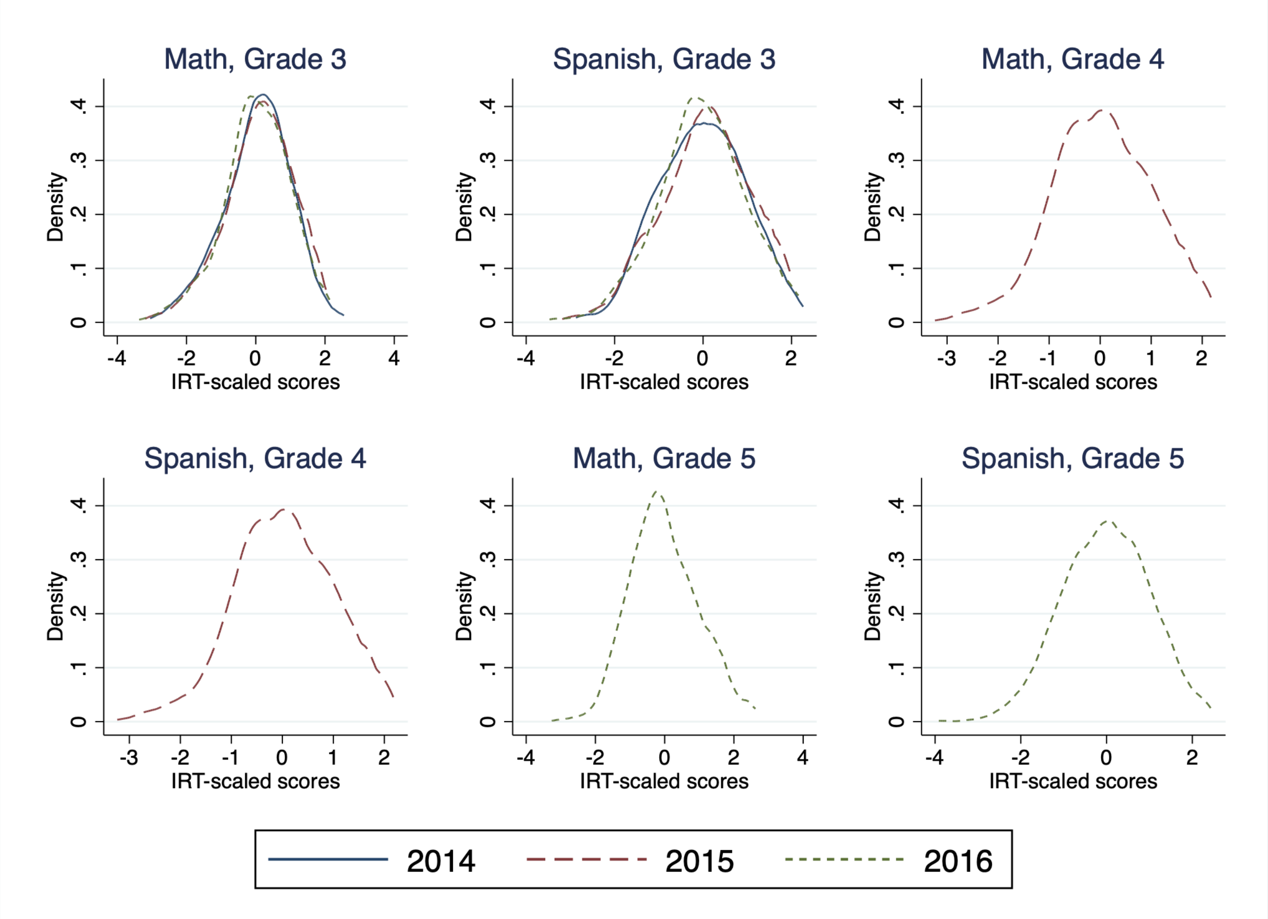
[32] We used the IRT program Stata 15 module (see Stata 2017). Our choice of a 2PL instead of a 3PL model was based partly on the sampling requirements for 3PL models discussed in Yen and Fitzpatrick (2006).

Figure B.1: Distribution of percent-correct scores on student assessments (2014-2016)



Notes: (1) This figure shows the distribution of percent-correct scores on the student assessments for this evaluation by subject and grade. (2) Each graph includes all students (i.e., control and treatment) assessed on a given year (see Table 1).

Figure B.2: Distribution of IRT-scaled scores on student assessments (2014-2016)



Notes: (1) This figure shows the distribution of IRT-scaled scores on the student assessments for this evaluation by subject and grade. (2) Each graph includes all students (i.e., control and treatment) assessed on a given year (see Table 1).

Table B.1: Comparison of distribution of items across content and cognitive domains in the assessments used for this study with national and international assessments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Assessment used in this study			National assessment		International assessments	
	(Salta)			(*Aprender*)		(TIMSS)	(PIRLS)
	Grade 3	Grade 4	Grade 5	Grade 3	Grade 6	Grade 4	Grade 4
Panel A. Math							
Content domains							
Number	60%	51%	43%	76%	50%	50%	
Geometry	17%	23%	20%	24%	36%	30%	
Measurement	23%	17%	26%				
Probability and statistics		9%	11%		14%	20%	
Cognitive domains							
Knowing	37%	43%	34%	26%	26%	40%	
Communicating	14%	14%	14%	19%	18%	40%	
Algorithms	26%	14%	17%	11%	7%		
Reasoning	23%	29%	34%	43%	49%	20%	
Panel B. Reading							
Content domains							
Narrative texts	41%	41%	41%				
Informative texts	41%	41%	41%				
Short texts	18%	18%	18%				
Cognitive domains							
Retrieving explicit information	29%	29%	29%	38%	23%		20%
Making straightforward inferences	29%	29%	29%				30%
Interpreting and integrating information	26%	26%	26%	49%	39%		30%
Evaluating and critiquing textual elements	17%	17%	17%	13%	38%		20%

Notes: (1) This table compares the percentage of items allotted to each content and cognitive domain in the assessment used for the present study (columns 1-2) with those for the national assessment *Aprender*, which assesses grades 3 and 6 (columns 3-4), and the international assessments Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Study (PIRLS), which assess grade 4 (columns 5-6). (2) The figures for the national assessment were obtained from SEE-MEDN (2018). The figures for the international assessments were obtained from IEA (2015, 2017). (3) The distribution of items across different types of texts is not reported for the national and international assessments. (4) In some cases, the terms used to describe content and cognitive domains vary across assessments (e.g., "probability and statistics" is categorized as "data" in TIMSS). (5) Figures that span multiple rows are reported as a single category for that assessment (e.g., "communicating" and "algorithms" is categorized as "analyzing" in TIMSS).