The Impact of Combining Performance-Management Tools and Training with Diagnostic Feedback in Public Schools: Experimental Evidence from Argentina*

Rafael de Hoyos† World Bank

Sharnic Djaker‡ New York University

Alejandro J. Ganimian§ New York University

Peter A. Holland¶ World Bank

January 16, 2024

Abstract

Providing principals with low-stakes information on their students' test scores has been shown to improve school management, instruction, and achievement in upper-middle income countries. We evaluate this approach by itself ("diagnostic feedback" or T1) and combined with tools and training ("performance management" or T2) through an experiment in 396 public primary schools in Salta, Argentina. After two years, T1 had null or adverse effects on on students' performance in school, but T2 reduced grade repetition (especially, among cohorts with more exposure), even a year after the interventions ended. We cannot rule out small-to-moderate effects on achievement. T2 also impacted teacher quality, student beliefs, bullying and discrimination, and extracurricular activities for high-exposure cohorts. Our results suggest that tools and training can effectively complement information in contexts of low principal capacity.

JEL codes: C93, I21, I22, I25

Keywords: diagnostic feedback, performance management, professional development, school management, student assessments, Argentina

*We gratefully acknowledge funding provided by the World Bank. 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 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.

‡Doctoral Candidate, NYU. E-mail: sharnic.djaker@nyu.edu.

§Assistant Professor of Applied Psychology and Economics, NYU. alejandro.ganimian@nyu.edu.

¶Lead Education Specialist, World Bank. E-mail: pholland@worldbank.org.

1 Introduction

There is mounting evidence that school-management practices matter for student achievement. Some studies have found that "school value-added" (i.e., student-achievement gains, adjusted for school and student characteristics) varies between and within schools over time, showing that some schools obtain better results than others serving comparable students (Bartanen, 2020; Branch, Hanushek, and Rivkin, 2012; Coelli and Green, 2012; Dhuey and Smith, 2014; Grissom, Kalogrides, and Loeb, 2015; Grissom, Egalite, and Lindsay, 2021; Laing et al., 2016). Others have identified management practices (e.g., teacher feedback, data-guided instruction, high expectations) that correlate with test scores, implying that differences in performance are at least partly due to what principals do and how they engage students, teachers, and parents (Angrist, Pathak, and Walters, 2013; Bloom et al., 2015; Crawfurd, 2017; Dobbie and Fryer, 2013; Leaver, Lemos, and Scur, 2019; Lemos, Muralidharan, and Scur, 2021; Tavares, 2015). And yet others have shown that schools that have adopted such practices are more effective, suggesting that their effect may be causal and that there may be interactions between them (Abdulkadiroğlu et al., 2016; Cohodes, Setren, and Walters, 2021; Dean and Jayachandran, 2019; Fryer, 2014; Gray-Lobe et al., 2022; Romero, Sandefur, and Sandholtz, 2020).

Despite the growing recognition of the importance of school management, there is still little evidence-based guidance on how to improve it in low- and middle-income countries (LMICs). According to a recent literature review (Anand et al., 2023), there are only 15 experimental or quasi-experimental evaluations of interventions targeting school principals in these contexts. The average effect across these studies was positive, but small: 0.03 standard deviations (SDs). Further, 43 of the 56 impacts on math and language test scores were statistically insignificant. Most interventions had null results on learning even when they shifted management practices, indicating that we have not yet even identified the right levers of change in these settings.

One approach with increasing evidence of effectiveness in LMICs is "diagnostic feedback." A study in the province of La Rioja, Argentina found that assessing students and informing their principals of the results raised test scores by 0.33 SDs in math and 0.36 SDs in language (de Hoyos, Ganimian, and Holland, 2021). A similar initiative in Mexico had positive results (de Hoyos, García-Moreno, and Patrinos, 2017). In fact, providing information on students' achievement was recently recognized as among the most cost-effective strategies to improve learning outcomes in LMICs by experts surveying the evidence (Akyeampong et al., 2023).[1]

The promising results of diagnostic feedback raise the question of whether school systems should go one step further and help principals act on the information that they receive. School-management capacity in LMICs is extremely low (Lemos and Scur, 2016), so principals

[1] An early experiment in India had led some to conclude that feedback had little "bite" if it was not tied to incentives (Muralidharan and Sundararaman, 2010). Yet, that initiative may have failed because many teachers in that context are frequently absent to school (Chaudhury et al., 2006; Muralidharan et al., 2017).

could benefit from additional assistance. Yet, if such 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 not be best positioned to offer such support.

In this paper, we present the results of an experiment testing the complementarity between providing principals with achievement data and offering them tools and training to act on it. We randomly assigned 100 public primary schools in the province of Salta, Argentina to: a "diagnostic-feedback" (T1) group, in which we tested students in math and reading and made results available to principals through reports (as in de Hoyos, Ganimian, and Holland, 2021); a "performance-management" (T2) group, in which we also offered principals and supervisors online tools and training to develop and monitor progress towards a school-improvement plan; or a control group, in which schools continued to operate exactly as prior to the study.

We report three main sets of results. First, we find that diagnostic feedback alone had an negative effect on students' performance in school, but it had a positive effect when combined with performance management. After two years, T1 schools had higher repetition and overage rates than control schools by 1.8 and 2.1 percentage points (pp.), respectively. Yet, T2 schools had lower repetition rates than both control and T1 schools by 1.5 and 3.5 pp., respectively. In fact, T2 schools still outperformed these two groups a year after both interventions ended.[2] The cohorts of students with greater exposure to the interventions saw even larger impacts.[3] Those who were assessed thrice saw larger increases in repetition and overage rates from T1; those whose teachers received reports twice saw larger reductions in repetition rates from T2. We interpret this pattern of results as suggesting that accountability (assessments) may lead schools to adopt counterproductive strategies, and that information (reports) may help schools focus their efforts, but only when they have the appropriate capacity (tools and training).

Second, despite the effects on school performance, neither intervention raised achievement. We compared the performance of grade 6 students on the national student assessment (the only primary-school grade for which all students are assessed every other year in Argentina) and found no statistically significant differences in any subject between experimental groups during the two years in which the interventions were implemented or a year after they ended. It could be, however, that effects on the these assessment were too small to be detected: for most of the years and subjects we considered, we could not rule out small-to-moderate effects. It is also possible that, by the time the study cohorts reached grade 6, effects had faded out. Lastly, on the year that the student cohorts with greater exposure to the interventions reached grade 6, the national assessment did not test math or reading (the two targeted subjects). Therefore, we do not see these results as conclusive evidence of null effects on achievement.

[2]We show our effects are not due to chance occurrence by comparing the school performance of the three experimental groups for a placebo student cohort that never received the intervention and finding no effects.

[3]Different grades were assessed on each year of the study: in 2014, only grade 3; in 2015, grades 3 and 4; and in 2016, grades 3 and 5. Thus, cohorts differ in both the number of times they were tested and the number of times their teachers received reports on their performance. We discuss the intervention in section 2.4.

Third, we used the surveys of students from the national assessment to identify potential mechanisms through which the interventions may have impacted their performance in school. However, the fact that they are all consistent in the experimental group in which they emerge and in their direction gives us some confidence that this was not the case. Among those with more intervention exposure, T1 and T2 students were more likely to report teachers explained topics and T2 students were more prone to say teachers listened to them. In that same cohort, T2 students were more likely than both their control and T1 counterparts to express that they fared well in the tested subjects and that they found them interesting. They were also less likely than control peers to report instances of bullying or discrimination. Finally, and consistent with the pattern of effects on school performance, T1 students were more prone than those in the control group to engage in non-academic activities (e.g., playing video games) and T2 students to participate in academic activities (e.g., learning a language).[4]

The present study contributes to research on improving school management in LMICs. We show that when diagnostic feedback is insufficient to prompt productive changes in school management (presumably, because principals do not know how to act on the information that they are given) performance-management tools and training can serve as a useful complement (potentially, by strengthening principals' capacity to implement a school improvement plan). Most prior evaluations have found that this approach had null effects on student outcomes (Aturupane et al., 2022; Blimpo, Evans, and Lahire, 2015; García-Moreno, Gertler, and Patrinos, 2019; Muralidharan and Singh, 2020). Others have previously found positive impacts (Lassibille et al., 2010; Tavares, 2015), but we also rule out impacts due to chance (by using a placebo cohort that never received the interventions), show exposure to the intervention relates to the magnitude of effects (by leveraging variability in exposure across student cohorts), and present evidence of impact on multiple plausible mechanisms (including teacher quality, student beliefs, bullying and discrimination, and extracurricular activities). To our knowledge, ours is among the most comprehensive accounts of the promise of this approach to date.

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

[4]We did not pre-register our analyses. Thus, it is possible that our statistically significant results are due to chance (given the number of statistical tests that we performed). We believe, however, that this is unlikely, because we find evidence of positive impacts on both outcomes (school performance) and mechanisms (teacher quality, student beliefs, bullying and discrimination, and extracurriculars and work) for the same experimental group (T2)—especially, among those cohorts with greater exposure to that intervention.

2 Experiment

2.1 Context

Schooling in Argentina is compulsory and free from age 4 until the end of secondary education. 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).[5] Argentina's 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).[6] The school year runs from February to December, but the start and end dates vary by province.

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, 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, 2019b). 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.[7]

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 results available by geographic region), and prohibiting the public disclosure of results at the

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

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

[7] 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 detailed discussion of these issues, see Ganimian, 2014; Ganimian, 2009; Ganimian, 2015b).

school, teacher, and student level in the 2006 National Education Law (see Ganimian, 2015a).[8] 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 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, 2019a).[9] Second, it was one of the few provinces with the political will to experiment with a sub-national assessment. At the time, it was endorsed by the governor and the education minister.

2.2 Sample

The sample for this study included 396 public primary schools in urban and semi-urban areas of Salta, which collectively served 147,379 students across all grades at the onset of our study.[10] We arrived at this sample as follows. We started with all 840 primary schools in the province in 2014 and we excluded all 87 private primary schools (because we were interested in improving school management in public schools), all 344 public primary schools in rural areas (because their geographic spread would have limited our capacity to implement the intervention),[11] and 13 schools with incomplete data (which we needed for random assignment of the interventions).

In-sample schools had more students enrolled, lower passing rates, and higher repetition rates than their out-of-sample counterparts (Table A.1 in Appendix A), both across grades 3 to 5 (the focus of our intervention, panel A) and in each of these grades (panels B-D), both when we compare them to all out-of-sample schools (column 5) and only to those in urban and semi-urban areas (column 6).[12] Yet, not all differences were statistically significant. In- and out-of-sample schools also differed on their overage and dropout rates, but the direction of differences depends on whether we compare all or just urban and semi-urban schools. Thus,

[8]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).

[9]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.

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

[11]Note, however, that while rural schools account for a large share of the number of public primary schools in the province (41%), they serve a small share of primary-school students (4.6%).

[12]Note, however, that out-of-sample schools are either private schools or public schools in rural areas.

our results are of interest to large, under-performing schools in non-rural areas in upper-middle income countries in Latin America facing similar school-management challenges.

2.3 Randomization

We randomly assigned schools to: a diagnostic-feedback (T1) group, in which we administered standardized tests of math and reading and made results available to principals through reports (50 schools); a performance-management (T2) group, in which we administered tests and delivered reports as in T1 and also offered principals and supervisors workshops and a dashboard to develop, implement, and monitor school-improvement plans (49 schools); or a business-as-usual control group (all remaining 297 schools in the sample). We stratified our randomization by geographic area (i.e., whether schools were in urban or semi-urban areas). This setup allows us to evaluate the causal effect of diagnostic feedback and performance management by themselves (by comparing control schools to T1 and T2 schools, respectively) and of the unique components in performance management (by comparing T1 and T2 schools).

Control, T1, and T2 schools were comparable at the start of the study on nearly all indicators of school performance (Table A.2). By chance, T1 schools had more students than control and T2 schools, so we account for total enrollment in all of our estimating equations.

2.4 Intervention

A domestic think tank (the *Centro de Estudios en Políticas Públicas* or CEPP) implemented the diagnostic-feedback (T1) and performance-management (T2) interventions for two years (2015 and 2016). As Table 1 shows, in T1 and T2 schools, CEPP assessed students at the end of each year and delivered reports based on those assessments at the start of the next year (e.g., they assessed students at the end of 2014 and delivered reports at the start of 2015). As the table also shows, in T2 schools, CEPP offered the workshops and dashboard both years.

Control schools proceeded as usual without any changes; in fact, they were "blind" to their experimental assignment. They did not participate in any rounds of data collection to prevent the possibility that such participation would trigger behavioral responses due to monitoring.

Both T1 and T2 schools had their students participate in multiple-choice assessments of math and reading at the end of each year. The assessments were designed by a consultant and they were based on national and provincial standards for primary education (the *Contenidos Básicos Comunes*, the *Núcleos de Aprendizaje Prioritarios*, and *Diseño Curricular de Salta*). The math assessments covered three topics (numeracy, geometry, and measurement) and four skills (recognition of problems, solving algorithms, problem solving, and communication). The reading assessments featured informative and narrative texts and covered four skills

(recognition of explicit/implicit information, textual analysis, and reflection about language). In 2014, we assessed grade 3; in 2015, grades 3 and 4; and in 2016, grades 3 and 5.[13]

T1 and T2 schools also received brief (10-page), user-friendly reports based on those tests. Each report had four sections: an introduction, describing the assessments and the number of participating students; an overview, displaying the school's mean score in each grade and subject and (from the second report onwards) changes in these scores from previous years;[14] a distributional analysis, plotting the typical score and ranges of scores for the school and other schools in the province for each grade and subject; and an item analysis, showing the share of students in the school who answered each item in each grade and subject correctly.[15]

Principals and supervisors in T2 schools were also invited to participate in 11 workshops (six in 2015 and five in 2016). The workshops helped principals to develop school-improvement plans (i.e., plans to improve one or more aspects of their school),[16] upload information on that plan to an online dashboard, and use the dashboard to monitor the plan's implementation. In 2015, the workshops also covered how to conduct classroom observations and give feedback, and in 2016, they also identified effective instructional practices in math and language.

Lastly, principals and supervisors in T2 schools also had access to an online dashboard. The dashboards had three main sections: one on each school's scores on the assessments, passing rates in math and reading, and student absences; another one on 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 a final one on the students' performance in school (e.g., passing, repetition, and dropout rates).[17]

We hypothesized that the T1 and T2 interventions could impact principals' school-management practices through three main channels: accountability, information, and capacity. If principals know how to improve but face weak incentives, the potential reputational costs from disappointing assessment results should prompt them to work harder (accountability).[18] If principals do not know what needs to improve, the reports should help them allocate time and resources towards the grades and subjects where they are most needed (information). And if they do not know how to improve, the workshops and dashboards should provide them with the skills they need to enact the changes that they wish to see in their schools (capacity).

[13]The assessments can be found at: `https://bit.ly/2vp1AoQ` (2014), `https://bit.ly/2TSEMqU` (2015), and `https://bit.ly/3d3gmTh` (2016).

[14]Scores were scaled using a two-parameter logistic item response theory model to link results across years. The mean score was set to be 50 and the standard deviation to be 10 for ease of interpretation.

[15]A template of the report, translated into English, can be found at: `https://bit.ly/2xCPKbq`.

[16]We do not have access to these plans, so we cannot synthesize their content.

[17]A template of the dashboard, translated into English, can be found at: `http://bit.ly/2cYBXOR`.

[18]One possible route of accountability was the school supervisors, who were aware that the principals had received reports on their students' achievement and could request to see them (as they do with other school records) and who also attended the workshops. Accountability could also stem from teachers and parents, but only if principals chose to share the reports with them, given that they were not required to do so.

In the results section, we leverage variation across cohorts in exposure to the assessments and reports over the course of the study to explore each of these potential mechanisms.

The interventions were implemented largely as intended (see columns 3-6 in Table 1). Nearly all T1 and T2 schools participated in the assessments on which the reports were based; the vast majority of T2 schools had at least one representative (principal or supervisor) attend the workshops (except for workshop 8, which was attended by about a fifth of T2 principals); and most T2 schools accessed the dashboard at least once during the first year of the study. Yet, T2 schools were less likely to attend workshops or use the dashboards in the second year. This decrease in take-up of these components suggests that effects for T2 are likely driven by the first year of implementation—especially for dashboards, which were rarely used in 2016.

3 Data

We obtained access to two types of administrative data: the annual census of schools, which collects information on students' performance in school (e.g., passing, repetition, and dropout rates); and the national student assessment, which tests students' knowledge on four subjects (depending on the year) and administers an accompanying surveys of students.

We see multiple advantages from using these administrative datasets for impact evaluation. First, they cover all schools in the system, allowing us to circumvent "site-selection bias" (i.e., systematic differences between the places where an intervention is first evaluated and those to which it is later expanded). Second, all schools are expected to participate in all the rounds of data collection that produce the information in these datasets, enabling us to minimize the risk of "differential attrition" (i.e., schools not participating due to their intervention status). Third, there is nothing link the interventions being evaluated and the data being collected, curtailing "social-desirability bias" (i.e., respondents acting based on perceived study goals). Lastly, using administrative data greatly reduces data-collection costs (it is essentially free). For all of these reasons, we have previously used these data for several impact evaluations (e.g., de Hoyos, Ganimian, and Holland, 2021; Ganimian, 2020; Ganimian and Freel, 2021).[19]

3.1 Annual census of schools

We have access to the annual census of schools (*Relevamiento Anual*) from 2014 (when students were first assessed for the school reports, allowing us to account for pre-intervention outcomes) to 2018 (two years after the interventions ended, enabling us to test for medium-term effects).

[19]We estimate the impact of the interventions on the same two sets of outcomes that we had included in our previous evaluation of similar interventions in La Rioja: students' performance in school and standardized tests. Yet, our analysis of potential mechanisms was largely determined by the questions included in the student surveys administered alongside the national assessment from 2016 to 2018, which vary across years.

We observe these data both for in- and out-of-sample schools, so we can compare schools in our sample to those in the rest of the system (to assess the external validity of our findings) and schools across experimental groups (to check randomization produced equivalent groups).

These data include the number of enrolled students, the passing rate (i.e., share of students who progressed to the next grade), the repetition rate (i.e., the share of students who had to repeat the grade), the overage rate (i.e., the share of students who are one or more years older than the expected age for their grade), and the dropout rate (i.e., the share of students who left their school without requesting a transfer to another school) at each grade in each school. Importantly, these data are not available at the student level, so we have limited statistical power to detect small effects on outcomes with very low or very high control-group means.

3.2 National student assessment

We also obtained access to the national student assessment (locally known as *Aprender*) from 2016 (the first year it was administered) to 2018 (two years after both interventions ended). We also observe these data for in- and out-of-sample schools and across experimental groups, but we do not use it to assess external validity or check for equivalence of expectation across groups because its first round of data collection occurred after the study had already began.

These data include students' scores on standardized tests of math and language for 2016 and 2018 (to assess the effect of the interventions on the two target subjects) and of natural and social sciences for 2017 (to explore potential spillover effects on non-target subjects).[20] They also include students' responses to surveys administered with each test, including items on teacher quality, student beliefs, bullying and discrimination, extracurriculars, and work, which we leverage to estimate impacts beyond achievement and identify mechanisms of impact. Importantly, the national assessment only covers all schools and students in grade 6,[21] so we cannot estimate the impact of the interventions on student achievement in the target grades. Yet, we leverage the fact that the cohorts of students who took these tests differed in their exposure to the components of the interventions (discussed in detail in section 5) to shed light on the relative importance of the intervention assessments and the associated school reports.

4 Empirical strategy

Our main equation for estimating the impacts of the interventions is:

[20] The subjects covered by the national assessment have varied across years. For an overview of the subjects assessed in each grade and year, see Ganimian, Pissinis, and Antonini, 2023.

[21] It assessed a sample of grade 3 students on math and language in 2016 and a sample of grade 4 students on writing in 2017, but we do not have access to these data at the student level.

$$Y_{igs} = \alpha_{r(s)} + X'_{igs}\gamma + \sum_{k=1}^{2} \beta_k T_k + \epsilon_{igs}, \tag{1}$$

where Y_{igs} is the outcome of interest for student i in grade g and school s on each year of the study; $r(s)$ is the randomization stratum of school s and $\alpha_{r(s)}$ is a stratum fixed effect; X'_{igs} is a vector of baseline covariates, including the school-level mean of the outcome of interest where available (to increase the precision of our estimates)[22] and the school-level total enrollment (to account for baseline imbalance); T_1 and T_2 are indicator variables for random assignment to the diagnostic-feedback and performance-management interventions, respectively; and ϵ_{igs} is an error term. The parameters β_1 and β_2 represent the intent-to-treat effect of the interventions. We estimate equation (1) by OLS regression. Standard errors are clustered at the school level.

5 Results

5.1 School performance

If we pool results across grades 3 to 5, by the end of the first year of the study, diagnostic-feedback (T1) and performance-management (T2) schools fared on par with control schools on school performance (i.e., passing, repetition, overage, and dropout rates; see Table 2, panel A). T1 schools had higher passing and lower dropout rates than control schools, but they also had higher repetition and overage rates, and none of the differences were statistically significant. T2 schools had more consistent results—higher passing rates and lower repetition, overage, and dropout rates—but once again, none of the differences reached statistical significance.

By the second year of the study, the divergent effects of T1 and T2 schools became clearer. T1 schools had (1.8 pp.) higher repetition rates and (2.1 pp.) higher overage rates than control schools (in both cases, $p < 0.05$; see panel B). These differences may seem small in absolute terms, but they represent 42% and 31% increases over the control means, respectively. T2 schools had (1.5 pp.) lower repetition rates than control schools ($p < 0.1$)—a 36% reduction. In fact, T2 schools outperformed T1 schools in both repetition rates (by 3.5 pp., $p < 0.01$) and dropout rates (by 3.6 pp., $p < 0.05$). These results suggest that, while the accountability and information components of T1 may lead schools to pursue unproductive strategies, when these are combined with the capacity add-ons of T2, schools can achieve meaningful improvements.

A year after the end of both interventions, most effects from the second year dissipated. There were no statistically significant differences between T1 and control schools, but T2 schools had lower repetition rates than control schools (by 1.4 pp., $p < 0.1$, representing a

[22] We cannot account for the student-level outcome of interest at baseline because, as discussed in section 3, the school-performance data is only reported at the school level and the national student assessment does not collect unique student identifiers.

40% reduction over the control group mean for that year) and T1 schools (by 2.1 pp, $p < 0.05$). As stated in section 3.1, because all of these indicators are reported at the school level, based on our 95% confidence intervals, we cannot rule out the possibility that T1 and T2 had negative and positive effects, respectively, that we lack statistical power to detect.

These short-lived impacts raise the possibility that the statistically significant differences that we observed in the second year of the interventions were merely due to chance occurrence. To explore this possibility, we estimate equation (1) for a cohort that was enrolled in the schools in our sample but that never received the interventions: students attending grade 4 in 2015. These students did not participate in any assessments (they were always a grade above the highest grade being assessed; see Table 1) and consequently, their teachers did not receive any reports, so we should not observe statistically significant differences across groups for them. This is precisely what we find: during the two years of the interventions and a year thereafter, the T1 and T2 groups fare similarly to the control group on school performance (Table A.3).[23]

In fact, the cohorts of students that seem to benefit most from the interventions are exactly those that either participated in more assessments or whose teachers received more reports. The students who were enrolled in grade 4 in 2015 participated in more assessments than any other cohort in our study: they were assessed in 2014 (when they were in grade 3), in 2015 (when they were in grade 4), and once again in 2016 (when they were in grade 6; see Table 1). If the interventions impacted school performance partly through the accountability channel discussed in section 2.4, we should observe statistically significant differences for this cohort. This is what we find: the adverse effects of T1 on repetition and overage rates by the second year (shown in Table 2) are particularly large for this student cohort (Table A.4, panel B). These results lend further support to our earlier assertion that accountability alone may lead some schools to pursue strategies that do not improve student outcomes.

Similarly, the students who were enrolled in grade 3 in 2015 had more teachers who received reports than any other cohort: their teachers received a report in 2015 (from grade 3 students who were assessed in 2014) and in 2016 (from grade 4 students assessed in 2015; see Table 1). If the interventions impacted school performance in part through the information mechanism discussed in section 2.4, we should also see statistically significant differences for this cohort. Once again, this is what we find: not just the adverse effects of T1 on repetition and overage rates, but also the desirable effects of T2 on passing and repetition rates (shown in Table 2) are larger for this cohort by the end of the second year of interventions (Table A.5, panel B). This pattern of results is consistent with our claim above that, while information alone may be insufficient to prompt improvements, it may do so when complemented with capacity-building.

[23]We only find one marginally statistically significant difference between the passing rates of control and T2 schools in 2015, which is likely due to chance given the number of comparisons that we are running.

5.2 Student achievement

If we estimate the impact of diagnostic feedback (T1) and performance management (T2) on the test scores of grade 6 students (the only grade for which the national assessment is census based; see section 3), we find no statistically significant effects from 2016 to 2018 (Table 3).[24]

In 2016, control students performed 0.17 and 0.20 SDs below the national mean in language and math, respectively (col. 1, panel A). T1 and T2 students outperformed them by a narrow margin (between 0.05 and 0.07 SDs, cols. 4-5). None of these differences were statistically significant, but based on the 95% confidence intervals around these estimates, we cannot rule out that these interventions had small-to-moderate positive or negative effects on test scores.[25] The differences between T1 and T2 are small (below 0.01 SDs) and statistically insignificant (col. 6). These null results are consistent with those for school performance for this cohort of students, which did not participate in assessments or receive reports (see Table A.3, panel B).

In 2017, control students performed 0.16 and 0.13 SDs under the national average in natural and social sciences (the only subjects tested by the national assessment on that year, which were not targeted by the intervention; col. 1, panel B). T1 students performed above and T2 students performed below their control peers (between 0.04 and 0.07 SDs, cols. 4-5). Again, differences are statistically insignificant, but they do not rule out moderate effects.[26] The differences between T1 and T2, while also statistically insignificant (col. 6), are consistent with the effects of the interventions on repetition rates for this cohort (Table A.4, panel B). Given that T1 increased repetition rates for grade 5 students in 2016, it could have prevented some low-achieving students from moving onto grade 6 in 2017, artificially boosting test scores.[27]

In 2018, control students performed on par with the rest of the country in language and 0.11 SDs below the mean in math (col. 1, panel C). Once again, T1 outperformed the control group (by 0.08 to 0.11 SDs; col. 4), but T2 performed on par with the control group (within 0.01 SDs; col. 5), so we can rule out moderate effects for the latter, but not for the former.[28] These effects are consistent with the differences between T1 and T2 on school performance for grade 5 students in 2017 (Table A.5, panel C), which suggest different types of students may have moved onto grade 6 in the T1 and T2 groups in 2018.[29]

[24] These assessments started in 2016, so we do not have data for prior years.

[25] These effects range from -0.12 to 0.24 SDs, depending on the subject and group.

[26] They range from -0.14 to 0.22 SDs, depending on the subject and group.

[27] We cannot test for this possibility directly, but impacts are slightly smaller (and still statistically insignificant) for 11-year-olds, the intended age for grade 6 (Table A.6), which is consistent with this possibility.

[28] For T1, the 95% confidence intervals include effects from -0.05 to 0.27 SDs, depending on the subject; for T2, they include effects from -0.10 to 0.11 SDs.

[29] We do not observe school performance for 2018, so we cannot verify this possible trend.

5.3 Teacher quality

The student surveys administered with the national assessment shed light on how diagnostic feedback (T1) and performance management (T2) changed students' experience at school.[30] If we estimate the effect of T1 and T2 on student perceptions of teachers, we only find statistically significant effects in 2017, the cohort with highest participation in assessments (Table 4).

In 2016, 81% of control students reported that teachers explain topics to students, 94% said that their teachers listen to students, 39% indicated that teachers make sure students understand the material, and 64% claimed that teachers often get upset (col. 1, panel A). The percentages of T1 and T2 students in these categories are nearly identical (cols. 2-3), which is not surprising given that this cohort of students did not receive any tests or reports.

In 2017, 67% of control students reported that teachers explain topics, 76% said that teachers listen to students, and 69% claimed that teachers praise good work (col. 1, panel B). T1 and T2 students were more likely to indicate that teachers explain topics by 4.7 or 6.8% ($p < 0.05$) and 6.5 pp. or 11% ($p < 0.01$), respectively (cols. 4-5). T2 students were also 3.6 pp. or 5.3% ($p < 0.01$) more likely to report that teachers listen to them.[31] The more robust effects on T2 are consistent with the impacts on school performance shown in Table 2.

In 2018, the surveys only had one question on students' perceptions of instructional quality: 68% of control students reported that teachers explain topics (col. 1, panel C). There were no statistically significant differences between any experimental group on this item (cols. 4-6).

5.4 Student beliefs

If we estimate the effect of diagnostic feedback (T1) and performance management (T2) on students' self-beliefs about academics, we also find that most statistically significant effects emerge in 2017, the cohort of students with highest participation in assessments (Table 5). We see these effects as further evidence of the impacts of T2. We cannot determine, however, whether they reflect actual improvements in students' ability or changes in their self-esteem.

In 2016, more than 93% of control students indicated that they understood and did well in math and language, leaving little room for impacts on these indicators (col. 1, panel A). T1 students were 1.6 pp. less likely to state they understand both math and language quickly ($p < 0.1$ and $p < 0.05$, respectively, col. 4). T2 students were 1.2 pp. or 2.7% more likely to claim that they understand language quickly ($p < 0.1$, col. 5), resulting in a 3.1 pp. difference between the T1 and T2 groups in the share of students in this category ($p < 0.01$, col. 6).

[30] As explained in section 3, these surveys were only administered to grade 6 students. The cohort of students who reached grade 6 by 2016 had not received no reports or tests, the one that reached it by 2017 had participated in four tests, and the one that reached it by 2018 had participated in one test (in grade 3 in 2015) and its teachers had received a report (in grade 4 in 2015), so results should be interpreted accordingly.

[31] As explained above, the questions included in the student surveys vary by year.

The percentage of control students who find math and language interesting was slightly lower (almost 80%), but nearly identical to those of T1 and T2 students.

In 2017, students were asked to rate their performance and interest in natural and social sciences, the two subjects on which they were evaluated in the national assessment that year, even if the tests and reports for both the T1 and T2 groups had focused on math and language. The shares of control students who reported understanding, doing well in, and finding these subjects interesting were markedly lower than those for math and language on the prior year (col. 1, panel B). T2 students were 3.1 pp. or 5.4% ($p < 0.05$) more likely to say that do well in social sciences, 2.7 pp. or 4.5% ($p < 0.05$) more prone to claim that they do well in natural sciences, and 3.3 pp. or 4.5% ($p < 0.01$) more likely to find that subject interesting (col. 5). In fact, we find statistically significant differences with T1 on the last two indicators (col. 6).

In 2018, the surveys only asked whether students performed well in math and language. The percentage of control students reporting doing well on these subjects were much closer to the 2017 figures for natural and social sciences than to the 2016 figures for math and language, but they were nearly identical across experimental groups.

5.5 Student bullying and discrimination

If we estimate the impact of diagnostic feedback (T1) and performance management (T2) on bullying and discrimination, we find some evidence in favor of T2 concentrated in the cohort that reached grade 6 in 2017, which is consistent with the results observed above (Table 6).

In 2016, students were only asked whether they get along with their classmates and whether peers mock or fight with each other. Roughly 73% of control students selected these options, which was approximately the same proportion of T1 and T2 students (cols 1-3, panel A).

In 2017 and 2018, students were asked many more questions on bullying and discrimination. In 2017, 91% of control students got along with their peers, but 22% reported that students with good grades are mocked and 27% that some are discriminated against (col. 1, panel B). T2 reduced the prevalence of these reports by 2.4 pp. or 13% and 3.1 pp. or 12%, respectively (both $p < 0.05$), in the latter case leading to an even larger difference with T1 (cols. 5-6). Other types of misbehavior were less prevalent and comparable across experimental groups, but the sign of the T2 coefficient is negative in nearly all of them, raising the possibility that T2 reduced their prevalence as well but we lacked statistical power to detect such effects.

In 2018, the incidence of bullying and discrimination among control students resembled that of 2017, but we find precisely estimated null effects for T1 and T2 (panel C, cols. 4-5), suggesting that gains on this front may have been cohort-dependent rather than broad-based.

5.6 Student extracurricular activities and work

The impacts of T1 and T2 on school-level outcomes could have reverberated on students' lives outside of school. Specifically, we wondered whether the mostly negative (positive) impacts of T1 (T2) could have decreased (increased) student engagement in extracurricular activities and increased (decreased) student involvement in either unpaid or paid labor outside of school. We see some evidence consistent with the former and almost no evidence of the latter.

Across 2016-2018, T1 tended to increase student engagement in non-academic activities (e.g., meeting up with friends, playing with a console or computer, or taking art lessons) and T2 increased participation in one academic activity (i.e., learning a language; see Table A.7). Yet, only the effect of T2 on students learning a language is statistically significant on two years: 2017 (3.3 pp. or 11%) and 2018 (2.9 pp. or 7.4%; both $p < 0.05$, col. 5, panels B-C). Consistent with this trend, T2 students were 2.3 pp. or 3.2% ($p < 0.1$) less likely to watch television than their T1 counterparts in 2017 (col. 6, panel B).

Across all three years, between one and two thirds of students reported working at home (e.g., helping their parents, taking care of a relative, or doing household chores; see Table A.8). Only between a tenth and a fifth of students worked outside the home (e.g., agricultural work). The coefficients of both T1 and T2 are estimated to be around zero for both types of work.

6 Conclusion

A rapidly growing body of research indicates that the way in which school principals run their schools has important consequences for how much students are able to learn. Yet, there is still little evidence on how to support principals to improve their school-management practices.

We build on prior studies demonstrating the positive effects of providing principals with low-stakes (i.e., diagnostic) information on how their students perform on standardized tests by presenting the results of an experiment in which we randomly assigned 100 public primary schools in Salta, Argentina to such information—alone ("diagnostic feedback") or combined with tools and training to act on it ("performance management")—or to a control condition. We find that information alone has null or adverse effects on students' performance in school, but when it is combined with tools and training, it reduces grade repetition (especially, among cohorts with more exposure) with respect to both the control and other treatment groups. These effects persist one year after the implementation of the interventions had concluded. Neither intervention impacted achievement, but we cannot rule out small-to-moderate effects. We also show that performance management impacted several potential mediators, including teacher quality, student beliefs, bullying and discrimination, and extracurricular activities.

Our study, when read alongside others that have also found positive effects from helping principals pursue action plans based on diagnostics (Lassibille et al., 2010; Tavares, 2015),

shows that it is possible to complement information provision to improve student outcomes. Most prior efforts to train principals to enact school-improvement plans have had null effects, both when implemented by themselves (Aturupane et al., 2022; García-Moreno, Gertler, and Patrinos, 2019) or in combination with some form of information provision (Blimpo, Evans, and Lahire, 2015; de Hoyos, Ganimian, and Holland, 2021; Muralidharan and Singh, 2020). We present the most compelling evaluation of the promise of this approach in LMICs to date.

A question we cannot answer is why this training works in some contexts and not others. Our study and others that have also found positive effects differ from those with null effects in many dimensions, so we cannot conclusively identify the reasons for their divergent results.[32] Yet, we hypothesize that they are partly explained by differences in the capacity of principals (i.e., the less principals know about how to act on the information they receive, the greater the margin for training to have an impact), the link between information and training (i.e., the more it addresses issues raised by information, the higher the chances that principals will act on them), and the relevance of the ensuing plans (i.e., the more they focus on binding constraints to learning, the more likely their implementation will improve benefit students). We believe future evaluations of similar interventions ought to collect data on these dimensions to advance our understanding of why some initiatives have been more successful than others.

[32]For example, Anand et al. (2023) note that most impact evaluations of school-management interventions in LMICs are under-powered, which explains why they produce null results.

References

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2016). "Charters without lotteries: Testing takeovers in New Orleans and Boston." *American Economic Review* 106 (7), pp. 1878–1920.

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.

Akyeampong, T., T. Andrabi, A. Banerjee, R. Banerji, S. Dynarksi, R. Glennerster, S. Grantham-McGregor, K. Muralidharan, B. Piper, S. Ruto, J. Saavedra, S. Schmelkes, and H. Yoshikawa (2023). *2023 cost-effective approaches to improve global learning. What does recent evidence tell us are "smart buys" for improving learning in low- and middle-income countries?* London, UK; Washington, DC; New York, NY: Foreign, Commonwealth & Development Office (FCDO), World Bank, United Nations International Children's Emergency Fund (UNICEF), United States Agency for International Development (USAID).

Anand, G., A. Atluri, L. Crawfurd, T. Pugatch, and K. Sheth (2023). "Improving school management in low and middle income countries: A systematic review." *Unpublished manuscript*.

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

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013). "Explaining charter school effectiveness." *American Economic Journal: Applied Economics* 5 (4), pp. 1–27.

Aturupane, Harsha, Paul Glewwe, Tomoko Utsumi, Suzanne Wisniewski, and Mari Shojo (2022). "The impact of Sri lanka's school-based management programme on teachers' pedagogical practices and student learning: evidence from a randomised controlled trial." *Journal of Development Effectiveness* 14 (4), pp. 285–305.

Bartanen, B. (2020). "Principal quality and student attendance." *Educational Researcher* 49 (2), pp. 101–113.

Blimpo, M. P., D. K. Evans, and N. Lahire (2015). "Parental human capital and effective school management: Evidence from the Gambia." (Policy Research Working Paper No. 7238). Washington, DC: The World Bank.

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

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).

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F. H. Rogers (2006). "Missing in action: Teacher and health worker absence in developing countries." *The Journal of Economic Perspectives* 20 (1), pp. 91–116.

Coelli, M. and D. A. Green (2012). "Leadership effects: School principals and student outcomes." *Economics of Education Review* 31 (1), pp. 92–109.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021). "Can successful schools replicate? Scaling up Boston's charter school sector." *American Economic Journal: Economic Policy* 13 (1), pp. 138–167.

Crawfurd, Lee (2017). "School management and public-private partnerships in Uganda." *Journal of African Economies* 26 (5), pp. 539–560.

de Hoyos, R., A. J. Ganimian, and P. A. Holland (2021). "Teaching *with* the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina." *World Bank Economic Research* 35 (2), pp. 499–520.

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, pp. 123–140.

Dean, J. and S. Jayachandran (2019). "The impact of early childhood education on child development in rural India." *Unpublished manuscript*. Karnataka, India: Abdul Latif Jameel Poverty Action Lab (J-PAL).

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), pp. 634–663.

DIEE (2020). "Anuario estadıstico 2019." Buenos Aires, Argentina: Dirección de Investigación y Estadıstica Educativa (DIEE).

Dobbie, W. and R. G. Fryer (2013). "Getting beneath the veil of effective schools: Evidence from New York City." *American Economic Journal: Applied Economics* 5 (4), pp. 28–60.

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. by A. V. Banerjee and E. Duflo. Vol. II. Oxford, UK: North Holland. Chap. 6: The personnel economics of the developing state.

Fryer, R. G. (2014). "Injecting charter school best practices into traditional public schools: Evidence from field experiments." *Quarterly Journal of Economics* 129 (3), pp. 1355–1407.

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).

— (2014). *Avances y desafıos pendientes: Informe sobre el desempeño de Argentina en el Tercer Estudio Regional Comparativo y Explicativo (TERCE) del 2013*. Ciudad Autónoma de Buenos Aires, Argentina: Proyecto Educar 2050.

— (2015a). *El termómetro educativo: Informe sobre el desempeño de Argentina en los Operativos Nacionales de Evaluación (ONE) 2005-2013*. Ciudad Autónoma de Buenos Aires, Argentina: Proyecto Educar 2050.

— (2015b). *Pistas hechas en Latinoamérica: ¿Qué hicieron los paıses, escuelas y estudiantes con mejor desempeño en el Tercer Estudio Regional Comparativo y Explicativo (TERCE)?* Ciudad Autónoma de Buenos Aires, Argentina: Red Latinoamericana por la Educación (Reduca) & Proyecto Educar 2050.

— (2020). "Growth mindset interventions at scale: Experimental evidence from Argentina." *Educational Evaluation and Policy Analysis* 42 (3), pp. 417–438.

Ganimian, A. J. and S. H. Freel (2021). "Can principal training improve school management? Short-term experimental evidence from Argentina." *Papeles de la Economıa Española* 166, pp. 67–83.

Ganimian, A. J., A. Pissinis, and S. Antonini (2023). *¿Qué aprendimos de Aprender? Informe sobre el desempeño de las 24 jurisdicciones argentinas en las evaluaciones nacionales*. Ciudad Autónoma de Buenos Aires, Argentina: Educar 2050.

García-Moreno, V. A., P. Gertler, and H. A. Patrinos (2019). "School-based management and learning outcomes: Experimental evidence from Colima, Mexico." (Policy Research Working Paper No. 8874). Washington, DC: The World Bank.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. W. Ozier (2022). "Can education be standardized? Evidence from Kenya." (Working Paper No. 2022-68). Chicago, IL: Becker Friedman Institute For Economics.

Grissom, J. A., A. J. Egalite, and C. A. Lindsay (2021). "How principals affect students and schools: A systematic synthesis of two decades of research." New York, NY: The Wallace Foundation.

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

Laing, D., S. Rivkin, J. Schiman, and J. Ward (2016). "Decentralized governance and the quality of school leadership." (NBER Working Paper No. 22061). Cambridge, MA: National Bureau of Economic Research (NBER).

Lassibille, G., J. Tan, C. Jesse, and T. Van Nguyen (2010). "Managing for results in primary education in Madagascar: Evaluating the impact of selected workflow interventions." *The World Bank Economic Review* 24 (2), pp. 303–329.

Leaver, C., R. Lemos, and D. Scur (2019). "Measuring and explaining management in schools: New approaches using public data." (Policy Research Working Paper No. 9053). Washington, DC: The World Bank.

Lemos, R., K. Muralidharan, and D. Scur (2021). "Personnel management and school productivity: Evidence from India." (NBER Working Paper No. 28336). National Bureau of Economic Research (NBER). Cambridge, MA.

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).

Muralidharan, K., J. Das, A. Holla, and A. Mohpal (2017). "The fiscal cost of weak governance: Evidence from teacher absence in India." *Journal of Public Economics* 145 (C), pp. 116–135.

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).

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

Romero, M., J. Sandefur, and W. A. Sandholtz (2020). "Outsourcing education: Experimental evidence from Liberia." *American Economic Review* 110 (2), pp. 364–400.

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

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

— (2019b). "Aprender 2018: Informe nacional de resultados, 6to año nivel primario." Ciudad Autónoma de Buenos Aires: Secretarıa de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

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

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).

Table 1: Timeline of the interventions

(1)	(2)	(3)	(4)	(5)	(6)
		School participation rates			
		Diagnostic	Performance management (T2)		
Month	Event	feedback (T1)	Overall	Principals	Supervisors
A. 2014					
Oct	Student assessments (grade 3)	96%	92%		
B. 2015					
Mar	Delivery of school reports				
	Workshop 1: Using data from dashboards		88%	70%	88%
Apr	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	Student assessments (grades 3 and 4)	92%	92%		
	Access to dashboards		78%		
C. 2016					
Apr	Delivery of school reports				
	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	Access to dashboards		8%		
	Student assessments (grades 3 and 5)	94%	92%		

Notes: 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). 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. The school year in Argentina runs from February to November (see section 2.1).

Table 2: ITT effect on students' performance in school, grades 3-5 (2015-2017)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2015						
Percentage of students who passed the grade	95.691	96.923	97.668	0.622	0.859	0.210
	(6.947)	(3.081)	(2.930)	[0.849]	[0.843]	[0.519]
Percentage of students who repeated the grade	2.965	2.872	2.661	0.117	-0.236	-0.327
	(4.585)	(3.675)	(4.166)	[0.591]	[0.607]	[0.801]
Percentage of students with overage	7.424	7.676	5.272	0.883	-1.338	-2.469**
	(9.298)	(8.203)	(6.320)	[0.897]	[0.899]	[0.997]
Percentage of students who dropped out of school	0.734	0.134	0.269	-0.300	-0.123	0.165
	(2.593)	(0.404)	(1.089)	[0.313]	[0.321]	[0.165]
B. 2016						
Percentage of students who passed the grade	94.677	95.833	97.513	0.414	1.641	1.577*
	(8.295)	(5.515)	(3.612)	[1.065]	[1.060]	[0.906]
Percentage of students who repeated the grade	4.256	5.626	2.655	1.823**	-1.511*	-3.526***
	(6.262)	(7.895)	(3.597)	[0.921]	[0.843]	[1.225]
Percentage of students with overage	6.829	7.971	4.600	2.087**	-1.319	-3.579**
	(10.424)	(10.959)	(8.263)	[1.026]	[1.013]	[1.392]
Percentage of students who dropped out of school	1.097	0.099	0.097	-0.615	-0.584	-0.003
	(4.463)	(0.457)	(0.455)	[0.596]	[0.599]	[0.094]
C. 2017						
Percentage of students who passed the grade	95.595	95.974	97.669	-0.293	1.198	1.492
	(8.341)	(5.816)	(3.891)	[1.155]	[1.143]	[0.931]
Percentage of students who repeated the grade	3.461	4.001	1.969	0.690	-1.376*	-2.062**
	(5.245)	(5.214)	(3.367)	[0.782]	[0.752]	[0.902]
Percentage of students with overage	6.443	6.341	5.041	1.316	-0.089	-1.616
	(10.207)	(9.296)	(11.020)	[1.042]	[1.120]	[1.429]
Percentage of students who dropped out of school	1.382	0.367	0.108	-0.689	-0.979	-0.277
	(5.240)	(1.462)	(0.534)	[0.740]	[0.739]	[0.212]
N (schools)	297	50	49	347	346	99

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). Panel A shows results for 2015, panel B for 2016, and panel C for 2017. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: ITT effect on student achievement, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Language (IRT-scaled score)	-0.172	-0.111	-0.096	0.057	0.069	0.008
	(0.899)	(0.991)	(0.914)	[0.093]	[0.061]	[0.116]
Math (IRT-scaled score)	-0.200	-0.152	-0.145	0.053	0.052	0.005
	(0.913)	(0.940)	(0.920)	[0.090]	[0.050]	[0.106]
N (schools)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Natural sciences (IRT-scaled score)	-0.159	-0.130	-0.200	0.037	-0.060	-0.073
	(0.918)	(0.960)	(0.942)	[0.077]	[0.043]	[0.090]
Social sciences (IRT-scaled score)	-0.134	-0.078	-0.157	0.065	-0.038	-0.063
	(0.952)	(1.013)	(0.932)	[0.082]	[0.046]	[0.087]
N (schools)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Language (IRT-scaled score)	0.003	0.082	0.024	0.078	-0.009	-0.055
	(0.819)	(0.834)	(0.792)	[0.069]	[0.046]	[0.080]
Math (IRT-scaled score)	-0.105	-0.010	-0.090	0.111	0.003	-0.086
	(0.967)	(0.971)	(0.925)	[0.082]	[0.052]	[0.087]
N (schools)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. All test scores are scaled using a two-parameter logistic Item Response Theory (IRT) model and standardized with respect to the national distribution. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: ITT effect on student perceptions of instruction quality, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Teachers explain topics to students	0.806	0.805	0.804	-0.000	-0.000	0.005
	(0.395)	(0.396)	(0.397)	[0.010]	[0.014]	[0.015]
Teachers listen to students	0.943	0.943	0.942	0.000	-0.001	0.000
	(0.232)	(0.232)	(0.234)	[0.005]	[0.006]	[0.009]
Teachers make sure students understand	0.388	0.373	0.388	-0.017	0.003	0.024
	(0.487)	(0.484)	(0.487)	[0.020]	[0.018]	[0.025]
Teachers get upset at students	0.644	0.663	0.650	0.016	0.006	0.006
	(0.479)	(0.473)	(0.477)	[0.017]	[0.020]	[0.025]
N (students)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Teachers explain topics to students	0.665	0.710	0.735	0.047**	0.065***	0.024
	(0.472)	(0.454)	(0.441)	[0.020]	[0.014]	[0.025]
Teachers listen to students	0.761	0.778	0.801	0.018	0.036***	0.023
	(0.426)	(0.416)	(0.400)	[0.015]	[0.010]	[0.017]
Teachers praise students when they do well	0.686	0.680	0.708	-0.003	0.018	0.030
	(0.464)	(0.467)	(0.455)	[0.016]	[0.016]	[0.022]
N (students)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Teachers explain topics to students	0.679	0.675	0.698	0.001	0.013	0.015
	(0.467)	(0.468)	(0.459)	[0.014]	[0.013]	[0.019]
N (students)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the student survey administered alongside the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. In all three years, students were asked how frequently they held a perception. In 2016, the scale ranged from 1 ("almost never") to 3 ("very often"); in 2017 and 2018, it ranged from 1 ("never") to 4 ("always"). For ease of interpretation, we coded the first set of responses as 0 if they were 1 and as 1 if they were 2 or 3, and we coded the second set as 0 if they were at or below 2 and as 1 if they were at or above 3, so that all means can be interpreted as the proportion of students who held a perception. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: ITT effect on student self-beliefs about academics, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Understands math quickly	0.930	0.913	0.926	-0.016*	-0.002	0.013
	(0.256)	(0.281)	(0.262)	[0.009]	[0.007]	[0.010]
Does well in math	0.968	0.969	0.966	0.002	-0.002	-0.006
	(0.176)	(0.172)	(0.182)	[0.005]	[0.005]	[0.007]
Thinks math is interesting	0.787	0.780	0.780	-0.005	-0.009	-0.005
	(0.410)	(0.415)	(0.414)	[0.015]	[0.015]	[0.020]
Understands language quickly	0.943	0.928	0.953	-0.016**	0.012*	0.031***
	(0.232)	(0.258)	(0.212)	[0.007]	[0.007]	[0.011]
Does well in language	0.970	0.970	0.967	-0.001	-0.003	-0.001
	(0.170)	(0.172)	(0.178)	[0.005]	[0.005]	[0.007]
Thinks language is interesting	0.796	0.807	0.787	0.013	-0.010	-0.023
	(0.403)	(0.394)	(0.410)	[0.012]	[0.014]	[0.017]
N (students)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Understands soc. sciences quickly	0.565	0.593	0.579	0.024	0.015	0.009
	(0.496)	(0.491)	(0.494)	[0.017]	[0.018]	[0.023]
Does well in soc. sciences	0.606	0.633	0.639	0.025	0.031**	0.026
	(0.489)	(0.482)	(0.480)	[0.018]	[0.014]	[0.020]
Thinks soc. sciences are interesting	0.721	0.741	0.751	0.020	0.030*	0.032
	(0.449)	(0.438)	(0.433)	[0.019]	[0.016]	[0.023]
Understands nat. sciences quickly	0.689	0.686	0.711	-0.005	0.019	0.038*
	(0.463)	(0.464)	(0.453)	[0.016]	[0.014]	[0.020]
Does well in nat. sciences	0.692	0.694	0.723	-0.000	0.027**	0.046**
	(0.462)	(0.461)	(0.448)	[0.015]	[0.014]	[0.018]
Thinks nat. sciences are interesting	0.778	0.782	0.813	0.003	0.033***	0.046***
	(0.416)	(0.413)	(0.390)	[0.014]	[0.011]	[0.017]
N (students)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Does well in math	0.627	0.615	0.621	-0.015	-0.003	0.003
	(0.484)	(0.487)	(0.485)	[0.017]	[0.017]	[0.024]
Does well in language	0.647	0.633	0.646	-0.015	-0.004	0.002
	(0.478)	(0.482)	(0.478)	[0.017]	[0.018]	[0.025]
N (students)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the student survey administered alongside the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. In all three years, students were asked how frequently they held a belief. In 2016, the scale ranged from 1 ("never") to 3 ("almost always"); in 2017 and 2018, it ranged from 1 ("never") to 4 ("always"). For ease of interpretation, we coded the first set of responses as 0 if they were 1 and as 1 if they were 2 or 3, and we coded the second set as 0 if they were at or below 2 and as 1 if they were at or above 3, so that all means can be interpreted as the proportion of students who held a belief. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: ITT effect on bullying and discrimination among students, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Gets along with peers	0.731	0.730	0.741	-0.003	0.006	0.003
	(0.444)	(0.444)	(0.438)	[0.015]	[0.015]	[0.021]
Peers mock or fight with each other	0.733	0.753	0.756	0.019	0.021	0.011
	(0.443)	(0.431)	(0.430)	[0.013]	[0.016]	[0.020]
N (students)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Gets along with peers	0.911	0.910	0.910	-0.002	-0.003	0.002
	(0.285)	(0.286)	(0.286)	[0.006]	[0.007]	[0.010]
Peers mock students who get good grades	0.221	0.223	0.193	0.004	-0.024**	-0.026
	(0.415)	(0.417)	(0.395)	[0.016]	[0.011]	[0.017]
Peers mock students who do poorly or repeat	0.255	0.250	0.231	-0.006	-0.020	-0.012
	(0.436)	(0.433)	(0.422)	[0.013]	[0.014]	[0.018]
Peers discriminate against certain groups	0.269	0.293	0.236	0.024	-0.031**	-0.056***
	(0.443)	(0.455)	(0.425)	[0.016]	[0.015]	[0.019]
Peers insult, threat, or hit students	0.267	0.257	0.251	-0.011	-0.014	-0.014
	(0.442)	(0.437)	(0.433)	[0.019]	[0.016]	[0.024]
Peers insult, threat, or hit teachers	0.072	0.076	0.065	0.007	-0.007	-0.017
	(0.259)	(0.265)	(0.246)	[0.010]	[0.008]	[0.015]
Peers steal	0.107	0.103	0.091	-0.003	-0.014	-0.012
	(0.309)	(0.304)	(0.288)	[0.011]	[0.010]	[0.013]
Peers damage school property	0.205	0.192	0.189	-0.017	-0.010	0.002
	(0.403)	(0.394)	(0.391)	[0.014]	[0.012]	[0.016]
Peers bully others on social media	0.141	0.138	0.144	-0.006	0.003	0.004
	(0.348)	(0.345)	(0.351)	[0.011]	[0.011]	[0.014]
N (students)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Gets along with peers	0.915	0.913	0.915	-0.002	-0.000	0.006
	(0.279)	(0.281)	(0.279)	[0.007]	[0.007]	[0.009]
Peers mock students who get good grades	0.186	0.182	0.189	-0.001	0.010	0.010
	(0.389)	(0.386)	(0.392)	[0.015]	[0.018]	[0.022]
Peers mock students who do poorly or repeat	0.201	0.183	0.209	-0.016	0.010	0.032
	(0.400)	(0.387)	(0.406)	[0.013]	[0.018]	[0.023]
Peers discriminate based on religion	0.078	0.069	0.078	-0.009	0.005	0.018
	(0.268)	(0.253)	(0.269)	[0.008]	[0.012]	[0.014]
Peers discriminate based on appearance	0.234	0.214	0.237	-0.020	0.002	0.020
	(0.423)	(0.411)	(0.426)	[0.016]	[0.021]	[0.027]
Peers discriminate based on disability	0.103	0.093	0.096	-0.009	-0.003	0.005
	(0.303)	(0.291)	(0.295)	[0.011]	[0.015]	[0.018]
Peers discriminate based on nationality	0.104	0.096	0.103	-0.012	0.002	0.014
	(0.305)	(0.294)	(0.305)	[0.011]	[0.016]	[0.020]
Peers damage school property	0.189	0.203	0.195	0.010	0.011	0.000
	(0.392)	(0.402)	(0.397)	[0.015]	[0.015]	[0.020]
Peers bully others on social media	0.135	0.132	0.138	-0.003	0.002	0.007
	(0.342)	(0.339)	(0.345)	[0.011]	[0.011]	[0.016]
N (students)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the student survey administered alongside the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. In all three years, students were asked how many students they got along with. In 2016, the scale ranged from 1 ("none") to 4 ("everyone"); in 2017 and 2018, it ranged from 1 ("none") to 5 ("everyone"). For ease of interpretation, we coded both sets of responses as 0 if they were at or below 2 and as 1 if they were at or above 3. Also in all three years, students were asked how frequently their peers engaged in a behavior. In 2016, the scale ranged from 1 ("almost never") to 3 ("many times"); in 2017 and 2018, it ranged from 1 ("never") to 4 ("always"). We coded the first set of responses as 0 if they were 1 and as 1 if they were 2 or 3, and we coded the second set as 0 if they were at or below 2 and as 1 if they were at or above 3, so that all means can be interpreted as the proportion of peers who engaged in a behavior. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A Additional figures and tables

Table A.1: Comparison between in- and out-of-sample schools on school performance (2014)

	(1) All schools	(2) Out-of-sample schools All	(3) Out-of-sample schools Non-rural	(4) In-sample schools	(5) Col.(4)-Col.(2)	(6) Col.(4)-Col.(3)
A. Grades 3-5						
Number of enrolled students	95.561	37.550	134.870	160.604	123.054***	25.734*
	(119.815)	(69.422)	(95.041)	(130.625)	[7.113]	[13.913]
Percentage of students who passed the grade	96.017	96.100	98.518	95.924	-0.176	-2.594***
	(8.427)	(10.214)	(2.636)	(5.852)	[0.585]	[0.612]
Percentage of students who repeated the grade	2.599	2.317	1.190	2.912	0.596	1.722***
	(6.686)	(8.327)	(2.572)	(4.160)	[0.464]	[0.438]
Percentage of students with overage	21.403	37.954	11.286	11.826	-26.128***	0.540
	(21.865)	(25.833)	(16.196)	(10.919)	[1.566]	[1.926]
Percentage of students who dropped out of school	0.720	0.753	0.087	0.683	-0.070	0.596**
	(3.176)	(3.471)	(0.391)	(2.816)	[0.220]	[0.288]
B. Grade 3						
Number of enrolled students	33.413	14.134	46.835	52.839	38.705***	6.004
	(39.767)	(24.428)	(31.489)	(42.706)	[2.478]	[4.620]
Percentage of students who passed the grade	95.511	95.678	98.791	95.345	-0.333	-3.446***
	(10.977)	(13.662)	(3.393)	(7.399)	[0.786]	[0.783]
Percentage of students who repeated the grade	2.990	2.327	0.908	3.659	1.332**	2.751***
	(8.486)	(10.130)	(3.088)	(6.361)	[0.604]	[0.665]
Percentage of students with overage	18.002	39.377	10.628	10.170	-29.207***	-0.458
	(21.288)	(26.265)	(13.863)	(11.795)	[2.018]	[2.984]
Percentage of students who dropped out of school	0.865	1.078	0.126	0.652	-0.426	0.526
	(5.620)	(7.187)	(0.770)	(3.399)	[0.402]	[0.353]
C. Grade 4						
Number of enrolled students	34.325	14.140	45.414	54.510	40.371***	9.096*
	(40.985)	(24.389)	(32.291)	(44.182)	[2.542]	[4.731]
Percentage of students who passed the grade	96.148	96.167	98.391	96.129	-0.038	-2.262***
	(11.160)	(13.628)	(2.973)	(7.942)	[0.794]	[0.825]
Percentage of students who repeated the grade	2.638	2.382	1.679	2.895	0.513	1.216**
	(7.580)	(9.467)	(4.329)	(5.030)	[0.540]	[0.551]
Percentage of students with overage	18.476	35.490	13.516	11.624	-23.866***	-1.892
	(21.269)	(28.804)	(22.111)	(11.776)	[2.000]	[2.502]
Percentage of students who dropped out of school	0.502	0.385	0.083	0.620	0.234	0.537*
	(3.215)	(3.407)	(0.733)	(3.008)	[0.228]	[0.309]
D. Grade 5						
Number of enrolled students	33.956	13.795	44.929	54.476	40.681***	9.547**
	(41.124)	(23.884)	(31.670)	(44.714)	[2.540]	[4.771]
Percentage of students who passed the grade	96.853	97.413	98.414	96.275	-1.138*	-2.140***
	(8.850)	(9.852)	(3.453)	(7.650)	[0.625]	[0.800]
Percentage of students who repeated the grade	2.088	1.948	1.002	2.230	0.282	1.228**
	(7.631)	(9.558)	(2.504)	(4.963)	[0.542]	[0.515]
Percentage of students with overage	20.929	39.606	15.775	13.089	-26.518***	-2.686
	(22.064)	(27.610)	(21.760)	(12.832)	[1.886]	[2.704]
Percentage of students who dropped out of school	0.801	0.814	0.037	0.787	-0.027	0.750**
	(4.539)	(5.379)	(0.364)	(3.472)	[0.321]	[0.355]
N (schools)	840	444	100	396	840	496

Notes: The table shows the means and standard deviations of all public primary schools in Salta (column 1), all out-of-sample schools (column 2) and those in urban and semi-urban areas (column 3), and in-sample schools (column 4). It also tests for differences between in-sample and all out-of-sample schools (column 5) and out-of-sample schools in urban and semi-urban areas (column 6). Panel A shows results across grades 3 to 5, and Panels B-D for each of those grades. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2: Balancing checks between experimental groups on school performance (2014)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. Grades 3-5						
Number of enrolled students	152.566	200.040	169.082	25.527*	-3.707	-29.185*
	(127.880)	(149.396)	(121.210)	[13.359]	[12.573]	[16.842]
Percentage of students who passed the grade	95.601	96.502	97.288	0.617	1.433	0.783
	(6.416)	(3.675)	(3.375)	[0.916]	[0.922]	[0.713]
Percentage of students who repeated the grade	2.932	2.781	2.928	-0.075	0.100	0.147
	(4.377)	(2.556)	(4.198)	[0.638]	[0.669]	[0.700]
Percentage of students with overage	12.291	9.998	11.034	-1.101	-0.086	0.940
	(11.309)	(8.576)	(10.741)	[1.476]	[1.504]	[1.596]
Percentage of students who dropped out of school	0.847	0.238	0.148	-0.498	-0.595	-0.090
	(3.217)	(0.789)	(0.426)	[0.453]	[0.456]	[0.128]
B. Grade 3						
Number of enrolled students	50.752	63.120	54.917	5.603	-2.831	-8.492
	(42.297)	(47.916)	(38.497)	[4.502]	[4.292]	[5.458]
Percentage of students who passed the grade	94.888	96.113	97.361	0.911	2.185*	1.242
	(8.087)	(4.705)	(4.237)	[1.176]	[1.182]	[0.912]
Percentage of students who repeated the grade	3.629	3.566	3.938	0.070	0.497	0.374
	(6.488)	(4.435)	(7.319)	[0.953]	[1.025]	[1.221]
Percentage of students with overage	10.119	10.237	10.411	1.745	2.571	1.322
	(10.854)	(13.058)	(15.795)	[1.800]	[2.002]	[2.619]
Percentage of students who dropped out of school	0.825	0.220	0.037	-0.517	-0.702	-0.182*
	(3.895)	(0.716)	(0.179)	[0.558]	[0.562]	[0.105]
C. Grade 4						
Number of enrolled students	51.372	69.673	58.306	10.196**	0.208	-9.874*
	(43.042)	(50.888)	(41.191)	[4.548]	[4.219]	[5.864]
Percentage of students who passed the grade	95.726	96.967	97.695	0.974	1.728	0.718
	(8.832)	(4.618)	(3.559)	[1.270]	[1.271]	[0.826]
Percentage of students who repeated the grade	3.084	2.749	1.894	-0.241	-1.092	-0.834
	(5.537)	(2.954)	(2.952)	[0.812]	[0.810]	[0.590]
Percentage of students with overage	12.588	8.590	9.739	-2.707	-2.120	0.726
	(12.930)	(6.500)	(8.713)	[1.830]	[1.769]	[1.450]
Percentage of students who dropped out of school	0.770	0.224	0.118	-0.436	-0.553	-0.108
	(3.412)	(1.326)	(0.622)	[0.485]	[0.485]	[0.208]
D. Grade 5						
Number of enrolled students	51.475	68.640	58.167	10.018**	-0.657	-10.779*
	(43.604)	(51.200)	(42.014)	[4.656]	[4.474]	[6.038]
Percentage of students who passed the grade	96.183	96.349	96.755	-0.129	0.317	0.416
	(8.342)	(4.955)	(5.246)	[1.203]	[1.220]	[1.024]
Percentage of students who repeated the grade	2.114	2.081	3.097	-0.017	1.040	1.019
	(5.277)	(2.629)	(4.810)	[0.766]	[0.814]	[0.778]
Percentage of students with overage	13.679	11.451	11.350	-1.143	-1.079	-0.017
	(13.855)	(8.347)	(10.070)	[1.920]	[1.962]	[1.618]
Percentage of students who dropped out of school	0.957	0.273	0.289	-0.540	-0.534	0.015
	(3.960)	(0.987)	(0.954)	[0.557]	[0.562]	[0.196]
N (schools)	297	50	49	347	346	99

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects (columns 4-6). Panel A shows results across grades 3 to 5, and Panels B-D for each of those grades. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: ITT effect on school performance for cohort with no reports or tests (2015-2017)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. Grade 5 in 2015						
Percentage of students who passed the grade	96.195	97.617	98.639	1.104	1.996*	0.930
	(8.821)	(4.591)	(2.515)	[1.196]	[1.180]	[0.742]
Percentage of students who repeated the grade	2.377	2.913	2.278	0.784	-0.329	-1.249
	(5.795)	(7.581)	(5.710)	[0.896]	[0.842]	[1.380]
Percentage of students with overage	9.623	7.641	6.989	-0.532	-1.100	-0.931
	(13.404)	(6.118)	(7.429)	[1.463]	[1.513]	[1.278]
Percentage of students who dropped out of school	0.842	0.095	0.076	-0.414	-0.419	-0.016
	(3.673)	(0.506)	(0.528)	[0.467]	[0.468]	[0.108]
B. Grade 6 in 2016						
Percentage of students who passed the grade	96.313	97.008	98.152	0.474	1.292	0.792
	(8.585)	(4.532)	(4.257)	[1.210]	[1.233]	[0.826]
Percentage of students who repeated the grade	3.068	4.587	2.292	1.591	-0.757	-2.371
	(6.664)	(9.005)	(4.641)	[1.044]	[0.927]	[1.452]
Percentage of students with overage	10.124	9.544	8.744	-1.276	-0.576	-0.204
	(13.744)	(11.686)	(9.137)	[1.573]	[1.532]	[1.674]
Percentage of students who dropped out of school	1.058	0.097	0.197	-0.791	-0.707	0.180*
	(4.793)	(0.546)	(0.693)	[0.683]	[0.695]	[0.094]
C. Grade 7 in 2017						
Percentage of students who passed the grade	98.133	98.666	99.445	-0.011	1.123	0.916**
	(5.990)	(2.765)	(1.335)	[0.816]	[0.801]	[0.440]
Percentage of students who repeated the grade	2.882	1.945	1.948	-0.734	-0.780	-0.318
	(7.572)	(3.743)	(6.239)	[1.044]	[1.065]	[0.893]
Percentage of students with overage	9.291	9.054	10.906	2.076	1.434	-0.903
	(11.854)	(11.624)	(14.454)	[1.650]	[1.701]	[2.408]
Percentage of students who dropped out of school	1.122	0.042	0.106	-0.794	-0.943	0.065
	(4.334)	(0.226)	(0.521)	[0.602]	[0.604]	[0.084]
N (schools)	297	50	49	347	346	99

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3) for students who were in grade 5 in 2015 (they were not assessed and their teachers did not received reports in any year). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). Panel A shows results across grades 3 to 5, and Panels B-D for each of those grades. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: ITT effect on school performance for cohort with most tests (2015-2017)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. Grade 4 in 2015						
Percentage of students who passed the grade	95.643	96.885	97.960	0.243	0.936	0.702
	(9.758)	(3.776)	(3.370)	[1.190]	[1.171]	[0.698]
Percentage of students who repeated the grade	3.787	2.854	2.978	-0.648	-0.271	0.394
	(6.928)	(3.909)	(5.027)	[0.987]	[0.996]	[0.918]
Percentage of students with overage	4.664	5.522	3.062	1.998*	-0.926	-3.091**
	(7.811)	(8.383)	(6.223)	[1.092]	[1.036]	[1.414]
Percentage of students who dropped out of school	0.723	0.243	0.333	-0.283	-0.174	0.131
	(3.027)	(1.010)	(1.590)	[0.425]	[0.433]	[0.251]
B. Grade 5 in 2016						
Percentage of students who passed the grade	95.777	96.429	98.599	0.552	2.393**	2.255**
	(8.517)	(5.478)	(2.948)	[1.066]	[1.062]	[0.873]
Percentage of students who repeated the grade	3.556	5.898	2.869	2.738**	-0.894	-4.058**
	(7.161)	(10.304)	(6.203)	[1.135]	[1.036]	[1.706]
Percentage of students with overage	8.928	10.254	5.997	3.217*	-2.800	-6.361***
	(13.845)	(13.839)	(12.624)	[1.740]	[1.702]	[2.116]
Percentage of students who dropped out of school	0.925	0.000	0.058	-0.537	-0.480	0.061
	(4.616)	(0.000)	(0.288)	[0.592]	[0.593]	[0.042]
C. Grade 6 in 2017						
Percentage of students who passed the grade	96.234	96.588	98.023	-0.142	1.094	1.146
	(9.074)	(6.155)	(5.278)	[1.289]	[1.293]	[1.015]
Percentage of students who repeated the grade	2.948	3.719	1.335	0.762	-1.590*	-2.266**
	(5.733)	(6.928)	(3.337)	[0.906]	[0.852]	[1.085]
Percentage of students with overage	10.688	8.608	8.231	-1.090	0.744	0.785
	(15.564)	(12.274)	(15.184)	[1.868]	[2.141]	[2.905]
Percentage of students who dropped out of school	1.606	0.259	1.011	-0.986	-0.219	0.773
	(6.372)	(1.447)	(4.637)	[0.906]	[0.943]	[0.711]
N (schools)	297	50	49	347	346	99

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3) for students who were in grade 4 in 2015 (they were assessed in 2014, 2015, and 2016 and their teachers received reports in 2015). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). Panel A shows results across grades 3 to 5, and Panels B-D for each of those grades. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: ITT effect on school performance for cohort with most reports (2015-2017)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. Grade 3 in 2015						
Percentage of students who passed the grade	95.260	96.220	96.303	0.462	-0.129	-0.571
	(8.137)	(4.595)	(4.957)	[1.121]	[1.141]	[0.929]
Percentage of students who repeated the grade	3.787	2.854	2.978	-0.914	-0.674	0.154
	(6.928)	(3.909)	(5.027)	[0.983]	[1.018]	[0.909]
Percentage of students with overage	4.664	5.522	3.062	0.423	-1.618	-2.214*
	(7.811)	(8.383)	(6.223)	[1.149]	[1.178]	[1.323]
Percentage of students who dropped out of school	0.723	0.243	0.333	-0.269	-0.126	0.192
	(3.027)	(1.010)	(1.590)	[0.421]	[0.435]	[0.277]
B. Grade 4 in 2016						
Percentage of students who passed the grade	94.506	96.770	98.346	1.627	3.155**	1.461*
	(10.086)	(5.442)	(2.416)	[1.440]	[1.427]	[0.865]
Percentage of students who repeated the grade	4.389	5.599	2.083	1.971*	-1.828*	-4.006***
	(7.611)	(8.947)	(3.322)	[1.186]	[1.083]	[1.365]
Percentage of students with overage	6.518	7.520	3.693	3.209**	-1.816	-4.713**
	(11.169)	(10.477)	(8.484)	[1.266]	[1.305]	[1.906]
Percentage of students who dropped out of school	1.077	0.014	0.050	-0.862	-0.790	0.050
	(5.616)	(0.097)	(0.242)	[0.806]	[0.809]	[0.037]
C. Grade 5 in 2017						
Percentage of students who passed the grade	96.084	95.452	98.398	-0.986	1.927	3.111***
	(9.412)	(7.119)	(3.417)	[1.322]	[1.303]	[1.018]
Percentage of students who repeated the grade	3.043	4.024	1.593	1.001	-1.517*	-2.573***
	(6.090)	(5.379)	(2.780)	[0.913]	[0.890]	[0.897]
Percentage of students with overage	7.597	8.118	5.810	2.673*	-1.806	-4.279**
	(12.612)	(10.246)	(14.150)	[1.366]	[1.455]	[1.983]
Percentage of students who dropped out of school	1.330	0.740	0.119	-0.286	-1.025	-0.887
	(5.570)	(3.925)	(0.682)	[0.826]	[0.805]	[0.571]
N (schools)	297	50	49	347	346	99

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3) for students who were in grade 3 in 2015 (they were assessed in 2015 and their teachers received reports in 2015 and 2016). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). Panel A shows results across grades 3 to 5, and Panels B-D for each of those grades. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: ITT effect on achievement of 11-year-olds, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Language (IRT-scaled score)	-0.104	-0.082	-0.071	0.010	0.029	0.014
	(0.882)	(0.991)	(0.906)	[0.102]	[0.070]	[0.131]
Math (IRT-scaled score)	-0.159	-0.119	-0.116	0.044	0.043	0.008
	(0.921)	(0.939)	(0.904)	[0.095]	[0.055]	[0.115]
N (schools)	5,385	1,180	1,010	6,565	6,395	2,190
B. 2017						
Natural sciences (IRT-scaled score)	-0.125	-0.053	-0.185	0.079	-0.079	-0.144
	(0.903)	(0.968)	(0.934)	[0.085]	[0.050]	[0.102]
Social sciences (IRT-scaled score)	-0.103	0.013	-0.136	0.127	-0.049	-0.142
	(0.921)	(1.039)	(0.917)	[0.089]	[0.050]	[0.094]
N (schools)	6,499	1,438	1,315	7,937	7,814	2,753
C. 2018						
Language (IRT-scaled score)	0.046	0.139	0.061	0.102	-0.015	-0.098
	(0.800)	(0.813)	(0.783)	[0.069]	[0.046]	[0.079]
Math (IRT-scaled score)	-0.080	0.033	-0.065	0.133	-0.003	-0.111
	(0.956)	(0.961)	(0.911)	[0.086]	[0.051]	[0.091]
N (schools)	6,926	1,486	1,359	8,412	8,285	2,845

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. All test scores are scaled using a two-parameter logistic Item Response Theory (IRT) model and standardized with respect to the national distribution. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: ITT effect on student extracurricular activities, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Played sports	0.857	0.856	0.852	-0.002	-0.004	-0.006
	(0.350)	(0.351)	(0.355)	[0.009]	[0.011]	[0.014]
Read a book	0.669	0.643	0.663	-0.020	-0.004	0.019
	(0.471)	(0.479)	(0.473)	[0.020]	[0.018]	[0.023]
Met up with friends	0.795	0.824	0.787	0.028**	-0.009	-0.029
	(0.404)	(0.381)	(0.410)	[0.012]	[0.014]	[0.018]
Learned a language	0.422	0.424	0.422	0.004	0.002	-0.006
	(0.494)	(0.494)	(0.494)	[0.015]	[0.016]	[0.020]
Went to a show or exhibit	0.570	0.592	0.583	0.017	0.003	-0.015
	(0.495)	(0.492)	(0.493)	[0.024]	[0.027]	[0.035]
N (students)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Played sports	0.763	0.771	0.769	0.012	0.008	-0.006
	(0.425)	(0.420)	(0.422)	[0.012]	[0.012]	[0.016]
Read a book	0.538	0.537	0.560	0.011	0.028	0.045
	(0.499)	(0.499)	(0.496)	[0.024]	[0.020]	[0.028]
Met up with friends	0.796	0.778	0.795	-0.018	0.001	0.014
	(0.403)	(0.416)	(0.404)	[0.014]	[0.015]	[0.019]
Learned a language	0.268	0.285	0.297	0.016	0.033**	0.021
	(0.443)	(0.452)	(0.457)	[0.014]	[0.015]	[0.019]
Went to a show or exhibit	0.332	0.346	0.319	0.016	-0.016	-0.031
	(0.471)	(0.476)	(0.466)	[0.014]	[0.014]	[0.019]
Watched TV	0.834	0.848	0.821	0.011	-0.013	-0.023*
	(0.372)	(0.359)	(0.384)	[0.008]	[0.010]	[0.013]
Played with console or computer	0.529	0.560	0.538	0.026*	0.004	-0.019
	(0.499)	(0.496)	(0.499)	[0.015]	[0.017]	[0.023]
Surfed the Internet	0.593	0.629	0.589	0.019	-0.017	-0.034
	(0.491)	(0.483)	(0.492)	[0.023]	[0.021]	[0.028]
Took painting, dancing or music lessons	0.302	0.324	0.308	0.023*	0.006	-0.000
	(0.459)	(0.468)	(0.462)	[0.013]	[0.013]	[0.017]
N (students)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Played sports	0.828	0.823	0.832	-0.003	0.004	0.002
	(0.377)	(0.382)	(0.374)	[0.010]	[0.010]	[0.015]
Read a book	0.639	0.630	0.638	-0.008	0.009	0.009
	(0.480)	(0.483)	(0.481)	[0.017]	[0.017]	[0.022]
Met up with friends	0.761	0.756	0.761	-0.002	-0.001	-0.000
	(0.427)	(0.429)	(0.426)	[0.011]	[0.015]	[0.018]
Learned a language	0.379	0.379	0.407	-0.002	0.029**	0.024
	(0.485)	(0.485)	(0.491)	[0.018]	[0.013]	[0.022]
Went to a show or exhibit	0.572	0.588	0.591	0.009	0.008	0.000
	(0.495)	(0.492)	(0.492)	[0.019]	[0.018]	[0.023]
N (students)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the student survey administered alongside the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. In all three years, students were asked whether they engaged in an activity or not, so all means can be interpreted as the proportion of students who engaged in an activity. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: ITT effect on student work, grade 6 (2016-2018)

	(1) Control	(2) Diagnostic feedback (T1)	(3) Performance mgmt. (T2)	(4) Col.(2)-Col.(1)	(5) Col.(3)-Col.(1)	(6) Col.(3)-Col.(2)
A. 2016						
Helped mother or father at work	0.615	0.612	0.595	0.002	-0.007	-0.013
	(0.487)	(0.487)	(0.491)	[0.016]	[0.015]	[0.020]
Took care of sibling or other relative	0.638	0.646	0.645	0.009	0.007	0.006
	(0.481)	(0.478)	(0.479)	[0.012]	[0.011]	[0.015]
Did household chores	0.742	0.751	0.732	0.012	-0.009	-0.019
	(0.437)	(0.432)	(0.443)	[0.012]	[0.016]	[0.019]
Did agricultural work	0.218	0.189	0.191	-0.018	-0.014	0.007
	(0.413)	(0.392)	(0.393)	[0.016]	[0.015]	[0.019]
Worked outside home	0.269	0.246	0.243	-0.019	-0.015	-0.000
	(0.443)	(0.431)	(0.429)	[0.014]	[0.016]	[0.018]
N (students)	12,517	2,670	2,264	15,187	14,781	4,934
B. 2017						
Helped mother or father at work	0.624	0.605	0.624	-0.012	0.010	0.019
	(0.484)	(0.489)	(0.485)	[0.019]	[0.017]	[0.024]
Took care of sibling or other relative	0.393	0.368	0.379	-0.024*	-0.010	0.004
	(0.488)	(0.482)	(0.485)	[0.014]	[0.014]	[0.019]
Did household chores	0.356	0.364	0.356	0.009	0.004	-0.005
	(0.479)	(0.481)	(0.479)	[0.014]	[0.017]	[0.023]
Did agricultural work	0.112	0.095	0.105	-0.005	0.002	0.006
	(0.315)	(0.293)	(0.307)	[0.010]	[0.011]	[0.014]
Worked outside home	0.123	0.114	0.118	-0.002	0.002	0.000
	(0.328)	(0.317)	(0.323)	[0.012]	[0.010]	[0.015]
N (students)	13,112	2,947	2,625	16,059	15,737	5,572
C. 2018						
Helped mother or father at work	0.588	0.572	0.567	-0.010	-0.006	0.003
	(0.492)	(0.495)	(0.496)	[0.018]	[0.019]	[0.026]
Took care of sibling or other relative	0.380	0.391	0.367	0.011	-0.005	-0.031
	(0.485)	(0.488)	(0.482)	[0.016]	[0.013]	[0.019]
Did household chores	0.359	0.335	0.338	-0.021	-0.017	0.001
	(0.480)	(0.472)	(0.473)	[0.016]	[0.016]	[0.023]
Did agricultural work	0.116	0.094	0.117	-0.011	0.011	0.022
	(0.320)	(0.292)	(0.322)	[0.010]	[0.012]	[0.015]
Worked outside home	0.113	0.099	0.113	-0.007	0.008	0.010
	(0.317)	(0.299)	(0.316)	[0.011]	[0.010]	[0.014]
N (students)	12,882	2,726	2,453	15,608	15,335	5,179

Notes: The table shows the means and standard deviations of schools in the control (column 1), diagnostic-feedback (column 2), and performance-management groups (column 3). It also tests for differences between groups, accounting for geographic area fixed effects and baseline enrollment (columns 4-6). All results are from the student survey administered alongside the national student assessment. The only primary-school grade for which that assessment is census-based (i.e., covers all students) is grade 6. Panel A shows results for 2016, panel B for 2017, and panel C for 2018. In 2016, students were asked whether they engaged in an activity or not. In 2017 and 2018, for some questions, students were asked whether they engaged in an activity or not; for other questions, they were asked how frequently they engaged in an activity, ranging from 1 ("never") to 4 ("always"). For ease of interpretation, we coded responses at or below 2 as 0 and responses at or above 3 as 1, so that all means can be interpreted as the proportion of students who engaged in an activity. * significant at 10%; ** significant at 5%; *** significant at 1%.